National Library
of Canada

Acquisitions and
Bibliographic Services Branch

395 Wellington Street
Ottawa, Ontario
K1A 0N4

Bibliothèque nationale
du Canada

Direction des acquisitions et
des services bibliographiques

395, rue Wellington
Ottawa (Ontario)
K1A 0N4

Your file Votre référence

Our file Notre référence

NOTICE

The quality of this microform is heavily dependent upon the quality of the original thesis submitted for microfilming. Every effort has been made to ensure the highest quality of reproduction possible.

If pages are missing, contact the university which granted the degree.

Some pages may have indistinct print especially if the original pages were typed with a poor typewriter ribbon or if the university sent us an inferior photocopy.

Reproduction in full or in part of this microform is governed by the Canadian Copyright Act, R.S.C. 1970, c. C-30, and subsequent amendments.

AVIS

La qualité de cette microforme dépend grandement de la qualité de la thèse soumise au microfilmage. Nous avons tout fait pour assurer une qualité supérieure de reproduction.

S'il manque des pages, veuillez communiquer avec l'université qui a conféré le grade.

La qualité d'impression de certaines pages peut laisser à désirer, surtout si les pages originales ont été dactylographiées à l'aide d'un ruban usé ou si l'université nous a fait parvenir une photocopie de qualité inférieure.

La reproduction, même partielle, de cette microforme est soumise à la Loi canadienne sur le droit d'auteur, SRC 1970, c. C-30, et ses amendements subséquents.

Canada

UNIVERSITY OF ALBERTA

Results on Conics and Quadrics in Computer Aided Geometric Design

BY

Wenping Wang ©

A thesis submitted to the Faculty of Graduate Studies and Research in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

DEPARTMENT OF COMPUTER SCIENCE

EDMONTON, ALBERTA

Fall 1992

National Library of Canada

Acquisitions and Bibliographic Services Branch

395 Wellington Street
Ottawa, Ontario
K1A 0N4

Bibliothèque nationale du Canada

Direction des acquisitions et des services bibliographiques

395, rue Wellington
Ottawa (Ontario)
K1A 0N4

Your file Votre référence

Our file Notre référence

The author has granted an irrevocable non-exclusive licence allowing the National Library of Canada to reproduce, loan, distribute or sell copies of his/her thesis by any means and in any form or format, making this thesis available to interested persons.

L'auteur a accordé une licence irrévocable et non exclusive permettant à la Bibliothèque nationale du Canada de reproduire, prêter, distribuer ou vendre des copies de sa thèse de quelque manière et sous quelque forme que ce soit pour mettre des exemplaires de cette thèse à la disposition des personnes intéressées.

The author retains ownership of the copyright in his/her thesis. Neither the thesis nor substantial extracts from it may be printed or otherwise reproduced without his/her permission.

L'auteur conserve la propriété du droit d'auteur qui protège sa thèse. Ni la thèse ni des extraits substantiels de celle-ci ne doivent être imprimés ou autrement reproduits sans son autorisation.

ISBN 0-315-77375-8

Canada

UNIVERSITY OF ALBERTA

RELEASE FORM

NAME OF AUTHOR: Wenping Wang

TITLE OF THESIS: Results on Conics and Quadrics in Computer Aided Geometric Design

DEGREE: Doctor of Philosophy

YEAR THIS DEGREE GRANTED: 1992

Permission is hereby granted to the University of Alberta Library to reproduce single copies of this thesis and to lend or sell such copies for private, scholarly or scientific research purposes only.

The author reserves other publication and other rights in association with the copyright in the thesis, and except as hereinbefore provided neither the thesis nor any substantial portion thereof may be printed or otherwise reproduced in any material form whatever without the author's prior written permission.

Permanent Address:
01A-9103, 112 street,
Edmonton, Alberta,
Canada T6G 2C5

Date: Oct. 9, 1992.

UNIVERSITY OF ALBERTA

FACULTY OF GRADUATE STUDIES AND RESEARCH

The undersigned certify that they have read, and recommend to the Faculty of Graduate Studies and Research for acceptance, a thesis entitled **Results on Conics and Quadrics in Computer Aided Geometric Design** submitted by **Wenping Wang** in partial fulfillment of the requirements for the degree of **Doctor of Philosophy.**

Barry Joe

Barry Joe - Supervisor

Stan Cabay

Stan Cabay

Joseph Culberson

Joseph Culberson

S D Riemenschneider

Sherman Riemenschneider

Anup Basu

Ronald Goldman - External Examiner

Date:

To my wife

Abstract

Rational curves and surfaces, which include conics and quadrics, have primary importance in representing geometric shapes in Computer Aided Geometric Design. Mainly conics and quadrics expressed in rational parametric forms are studied in this thesis. The main contributions of this thesis are results on:

- Implementation issues for a simple method for drawing conic sections using a difference equation.
- Classification by reparameterization of faithful rational quadratic parameterizations of quadrics in E^3.
- Rational quadratic spline curve interpolation on a nondegenerate quadric in E^d, $d \geq 3$.
- Biarc interpolation on a sphere in E^d, $d \geq 3$, which leads to a simple method for interpolating orientations in E^3.

While the common subjects of these studies are conics and quadrics, they are basically independent of each other. Therefore this thesis is a collection of research on similar topics.

Acknowledgements

I am grateful to my supervisor, Professor Barry Joe, for his insightful guidance, many helpful discussions, countless readings of the thesis draft, financial assistance, and encouragement during this research. Without all these supports, this thesis would have been impossible.

I wish to thank Professors Stan Cabay, Joe Culberson, Sherman Riemenschneider, and Anup Basu for serving on my examining committee. This thesis has benefited greatly from their careful review.

I am grateful to my external examiner, Professor Ronald Goldman from Rice University, for his careful and critical reading of the thesis draft and his invaluable advice. I especially thank him for pointing out an earlier mistake in Chapter 3 by bringing to my attention the existence of unfaithful quadratic parameterizations of a quadric cone.

I am indebted to Professor Jiaye Wang, from Shandong University, for introducing me to the field of computer graphics, and many discussions about the contents of Chapter 2. I also wish to thank Mr. Anwei Liu for helpful discussions about many problems in this thesis.

Finally, I wish to acknowledge the Department of Computing Science at University of Alberta for providing equipment and the financial support during the research for this thesis.

Contents

1 **Introduction** 1

1.1 Rendering of conics . . . 2

1.2 Parametric representations and reparameterizations of quadrics . . . 3

1.3 Interpolation on quadrics by rational quadratic curves . . . 3

1.4 Spherical biarcs . . . 4

2 **Difference Method for Rendering Conics** 6

2.1 Introduction . . . 6

2.2 Difference method . . . 9

2.2.1 Conics with center at origin . . . 9

2.2.2 General conics . . . 11

2.2.3 Equivalence with matrix iteration method . . . 14

2.2.4 Approximation property and efficiency . . . 15

2.3 Drawing conic arcs to desired tolerance . . . 16

2.4 Conversion from parametric and implicit forms of conics . . . 23

2.4.1 Conics as rational quadratic curves . . . 24

2.4.2 A new parameterization of elliptic arcs . . . 29

2.4.3 Conics in implicit form 34

2.5 More properties of the difference method 37

2.5.1 A best approximation property 37

2.5.2 A projective property 40

2.5.3 On generalization 43

2.6 Summary 48

3 Parametric Representations and Reparameterizations of Quadrics 50

3.1 Preliminaries 50

3.2 Reparameterization 52

3.2.1 Rational linear reparameterization 54

3.2.2 General reparameterization 56

3.3 Rational parameterizations of quadrics 57

3.3.1 Parameterizations by projection 57

3.3.2 General parameterizations 61

3.4 Bézier representation of a triangular patch on a quadric 65

3.4.1 General triangular patches 66

3.4.2 Quadratic Bézier surfaces on a quadric 67

4 Interpolation on Quadric Surfaces by Rational Quadratic Spline Curves 69

4.1 Introduction 70

4.1.1 Problems 70

4.1.2 Preliminaries 71

4.2 Point interpolation on a quadric 75

4.2.1 Local representation 75

4.2.2 Construction of interpolating splines 81

4.2.3 Closed interpolating splines 84

4.3 Biarc interpolation on a quadric 86

4.3.1 Biarcs with positive weights 87

4.3.2 General biarcs 94

4.3.3 Properties of biarcs with positive weights 97

4.4 Point interpolation using biarcs 99

4.5 Summary 101

5 Theory of Spherical Biarc Interpolation **103**

5.1 Introduction 104

5.2 Preliminaries 105

5.3 Existence of spherical biarc interpolants 108

5.3.1 Existence for regular data 108

5.3.2 Existence for singular data 116

5.4 Properties of spherical biarcs 119

5.4.1 Fundamental properties 119

5.4.2 Perturbation of singular data 127

5.5 Summary 134

6 Shape Control and Applications of Spherical Biarcs **135**

6.1 Criteria and choices for shape control 135

6.1.1 Criteria 136

6.1.2 Equal chord biarc 138

6.1.3 Minimal twist biarc 141

6.2 Evaluation of three choices 146

6.3 Interpolation of a point sequence 155

6.4 Interpolation of unit quaternions for orientation modeling 158

6.4.1 About quaternions 158

6.4.2 Orientation interpolation problem 159

6.4.3 Spherical biarc solution 161

6.4.4 Comparisons 163

6.5 Summary 164

7 Concluding Remarks **165**

7.1 Summary 165

7.2 Open problems 167

List of Tables

Table Page

6.2.1 $mean\_R_i$ and $stdv\_R_i$ 149

6.2.2 min_R_i 149

6.2.3 max_ad_i and min_rd_i 150

6.2.4 $mean\_R_i$ and $stdv\_R_i$ with $|X_0X_1| > 0.4$ 152

6.2.5 min_R_i with $|X_0X_1| > 0.4$ 152

6.2.6 max_ad_i and min_rd_i with $|X_0X_1| > 0.4$ 153

6.2.7 $mean\_R_i$ and $stdv\_R_i$ with $|X_0X_1| < 0.1$ 153

6.2.8 min_R_i with $|X_0X_1| < 0.1$ 154

6.2.9 max_ad_i and min_rd_i with $|X_0X_1| < 0.1$ 154

List of Figures

Figure Page

2.2.1 Four ellipses 13

2.3.1 A hyperbolic arc 21

2.4.1 A rational quadratic Bézier curve with $\kappa = 25/16$ 26

2.5.1 A counterexample 41

2.5.2 The cubic curve $P(t)$ and its projection 45

4.2.1 Local interpolation on hyperboloid of one sheet 80

4.2.2 Control points 81

4.2.3 Data on hyperboloid of one sheet 84

4.2.4 Interpolations on a sphere 85

4.3.1 A spherical biarc 87

4.3.2 Singular data 90

4.3.3 Biarc interpolations on a sphere 95

4.3.4 Biarc interpolations for singular data 97

4.4.1 Spline interpolation using biarcs 100

5.2.1 Two cases indicated in Lemma 5.2.1 107

5.3.1 Regular data and singular data 111

5.3.2 The six cases in the proof of Theorem 5.3.4 115

5.3.3 A family of biarc on S^2 interpolating regular data 117

5.4.1 The hyperbola $F(k_0, k_1) = 0$ 124

5.4.2 The section of the sphere S intersected by the plane KNP_0 130

6.1.1 Comparison of two biarcs 138

6.2.1 151

6.2.2 Interpolations by the three biarcs 156

6.3.1 Interpolation of data on S^2 by Algorithm 6.3.1 157

6.4.1 Orientation interpolation 162

Chapter 1

Introduction

Computer Aided Geometric Design (CAGD) is an area of computer science that concerns itself with representing, manipulating, and visualizing geometric shapes. It has applications in computer graphics, as well as in practical fields such as the shipbuilding and automobile industries.

Rational curves and surfaces have primary importance in representing geometric shapes in CAGD. At the inception of CAGD, mainly polynomial curves and surfaces were used. Later, these were extended to rational curves and surfaces when it was realized that some of the most useful geometric entities, such as conics and quadrics can be exactly represented as rational curves or surfaces. Also, rational curves and surfaces are closed under projective transformations. In addition, in free form curve and surface design, rational curves and surfaces provide more flexibility than their polynomial counterparts. Currently rational curves and surfaces are one of the two main techniques for shape representation in CAGD; the other one is the implicit representation where a curve or surface is represented as the zero set of an algebraic equation $f(x,y) = 0$ or $f(x,y,z) = 0$, called an *implicit curve or surface.*

In this thesis we will mainly consider rational curves and surfaces, especially those of degree two. They include all conics and quadrics. It is well known from classical algebraic geometry that a rational curve or surface can always be represented as an implicit curve or surface. The problems we will study are (1) rendering conics using a difference equation; (2) parametric representation and reparameterization of quadrics; (3) interpolation on quadric surfaces using rational curves; and (4) spherical biarc interpolation.

The difference method studied in problem (1) is an important contribution of this thesis.

In view of its simplicity and its nice properties, through further research it may become a fundamental algorithm in computer graphics for drawing conics. Topics (2) and (3) are mainly theoretical investigations. The last topic, the theory of spherical biarc interpolation, is the main contribution of this thesis. This work yields a simple solution to the smooth curve interpolation problem on a sphere. It can be used to solve the orientation interpolation problem in 3-D space, which arises in computer animation and computer graphics.

While the common subjects of these studies are conics or quadrics, they are basically independent of each other. Therefore this thesis is a collection of research on similar topics.

1.1 Rendering of conics

Conics are rational quadratic curves. They are widely used in computer graphics and geometric modeling, and devising efficient rendering algorithms is an important research topic. In computer graphics rendering conics involves computing points on or close to the conic. The method we will discuss uses a difference equation to compute an inscribed polygon of a conic.

The two most commonly used methods for rendering conics are the matrix iteration method based on the recurrence $P_{i+1} = MP_i$, $i \geq 0$ [Smi72, Pav83], and the forward differencing method based on the rational parameterization of conics [BBB87]. The former generates point sequences of better quality while the latter is more efficient. The method we will study is a generalization of forward differencing. It uses the difference equation

$$P_{i+3} = (2k+1)(P_{i+2} - P_{i+1}) + P_i, \quad i \geq 0,$$

where P_0, P_1, P_2 are appropriately chosen initial points and k is a constant whose range depends on the type of the conic to be drawn. This method will be called the *difference method.* The difference method is more efficient than the forward differencing method, and generates the same point sequences as the matrix iteration method. In Chapter 2 the theory and implementation issues of the difference method will be studied in detail. This extends the work in [WaW89a, WaW89b].

1.2 Parametric representations and reparameterizations of quadrics

Reparameterization of a parametric curve or surface involves changing the parameter(s) of the curve or surface to other related parameter(s). Here we study rational reparameterizations of rational curves and surfaces, i.e. the function(s) relating the new and old parameters are rational functions. Sometimes the term *reparameterization* is also used to refer to the curve or surface with the new parameter(s). Reparameterization is a useful technique. For instance, it can be used to simplify the representation of a curve segment or surface patch in the Bézier form.

In Chapter 3 we will mainly consider linear and nonlinear rational reparameterization of rational surfaces, especially in the setting of rational surfaces which are quadric surfaces, and show that the situation is more complex than that for rational curve. It is known that any quadric can be represented as a rational quadratic surface [Som51, p.192]. We will study rational parameterizations of quadrics and the relationship between different rational parameterizations of a quadric surface in E^3. It is shown that (1) any rational quadratic parameterization of a quadric $S \subset E^3$ is associated with a unique point on S; (2) conversely, associated with any point on S there exists a family of rational quadratic parameterizations of S, which are rational linear reparameterizations of each other; (3) any two different parameterizations associated with different points on S are rational quadratic reparameterizations of each other; (4) any rational quadratic parameterization of quadric S has a rational linear inversion formula.

The above theory will then be used to obtain a rational representation of a triangular patch on a quadric with rational boundary curves. It is shown that any triangular patch on a quadric with rational boundary curves of degree n can be represented as a rational triangular Bézier patch of degree $2n$. In particular, if there are three planes each containing a boundary (a curved side) of the triangular patch, and the three planes intersect at a point on S, then the patch can be represented as a rational quadratic Bézier patch. This extends the results in [JoW90].

1.3 Interpolation on quadrics by rational quadratic curves

Motivated by the point interpolation problem in the unit quaternion space, which arises in computer animation [Sho85], in Chapter 4 we will study the generalized problem of interpolating a point sequence on a proper quadric S using rational quadratic spline curves, with the curve lying on S.

We first consider using one piece of a rational quadratic curve (conic arc) between two consecutive interpolated points. Several results on the existence and construction of such splines are proved. For instance, it is shown that (1) such a spline always exists on a sphere $S \subset E^d$, $d \geq 3$, and there are ∞^1 such splines, where ∞^k denotes that the number of free parameters is k; and (2) on a general quadric such a spline may not exist. A necessary condition for the existence of such an interpolating spline is that all line segments connecting consecutive interpolated points lie on the same side of the quadric.

A spline curve obtained in the above manner is not locally controllable even if it exists. So we further consider using two smoothly joining rational quadratic Bézier curves, called a *conic biarc*, between two consecutive interpolated points. In this setting a biarc is used to interpolate two points on a quadric and two tangent directions defined at the two points; this is called *Hermite interpolation*. To guarantee that these rational quadratic Bézier curves are continuous, we consider only quadratic Bézier curves with the two end weights being 1 and the middle weight being positive. The resulting conic biarc is called a *biarc with positive weights*. Several results on the existence and construction of biarcs with positive weights solving a Hermite interpolation problem on a quadric are derived.

1.4 Spherical biarcs

A spherical biarc contains two smoothly joining circular arcs on a sphere. In Chapter 5, we will consider using a spherical biarc to solve the Hermite interpolation problem on a sphere, i.e. to interpolate two points and two associated tangents on a sphere $S^{d-1} \subset E^d$, $d \geq 3$. This is a continuation of the discussion in Chapter 4, with the quadric specialized to a sphere in E^d.

Biarcs, especially plane biarcs, have been studied extensively in CAGD [Bez72, Sab76, SuL89]. Recently, space biarcs have found applications in surface modeling using cyclides [NuM88]. Biarcs are favored because of their obvious advantages: they can be easily represented, efficiently stored, and easily processed. We first consider the existence and construction of all spherical biarcs interpolating *data* comprising two given points and their associated tangent directions on a sphere in E^d, $d \geq 3$. It is shown that (1) there exists an essential distinction between two kinds of data, to be called *regular* and *singular data*; (2) there exist ∞^1 spherical biarcs interpolating any given regular data D, with the joints of these biarcs forming a circle on S, where a joint of a biarc is the point where the two arcs of the biarc meet; (3) there exist ∞^{d-2} spherical biarcs interpolating any given singular data, with the joints forming a $d-2$ dimensional sphere on S. Several more properties of spherical biarcs are derived in Chapter 5.

In Chapter 6, the shape control of spherical biarcs is discussed, i.e. how to choose a biarc with a fair shape among infinitely possibilities. It is shown that the biarc with the chords of its two arcs having equal length is a reasonably good choice. Using this shape control scheme, an algorithm is proposed for interpolating a sequence of points on a sphere using spherical biarcs.

In Chapter 6, we also address the application of spherical biarcs to interpolating object orientations in E^3. The idea is to use a spherical biarc spline to interpolate a point sequence in the unit quaternion space. Some examples of this application are illustrated. The point interpolation problem in the unit quaternion space has been solved in the literature using various approaches [Sho85, Sho87, Ple89, GeR91]. In view of efficiency and simplicity, we conclude that the spherical biarc is the best among all the existing G^1 schemes.

Chapter 2

Difference Method for Rendering Conics

In graphics, drawing a curve involves computing a sequence of points on or near the curve. In this chapter we study the computation of point sequences on a conic by using a difference equation, called the *difference method*. Applications and implementation issues of the difference method are addressed.

2.1 Introduction

Conic sections are well-studied curves, both as quadratic algebraic curves and rational quadratic parametric curves. They are widely used in geometric modeling and computer graphics. Although in practice various other curve schemes have been found very useful, in many cases the relatively simple conic arcs can also fulfill the demand. For example, in [Boo79, Far89, Pav83, Pra85] conics are used in a composite conic spline to solve interpolation and approximation problems. So it is still of interest to design efficient algorithms for rendering conics.

There are a number of algorithms proposed in the literature of computer graphics to render conics. Basically these algorithms fall into two categories: one constructs the pixel representation of the curve, that is, pixels on a display device which approximate the curve are generated; the other computes a sequence of connected line segments to represent the curve and then the line segments are rendered with some efficient line-drawing algorithm. The method we shall discuss belongs to the second class. Specifically, we shall show how to compute an inscribed polygon approximating a conic arc using a difference equation.

There are several other ways of generating inscribed polygons for conics, usually based on different representations of conics. We will mention the two most efficient and the most used ones below, with some of their properties listed for later reference. In order to be specific in efficiency measurement, and for simplicity of analysis, we shall mainly discuss conics in the plane E^2.

The most popular method for rendering conics in applications is the forward differencing method based on the rational parameterization of conics [BBB87, p. 400]. This method essentially uses the difference equation

$$P_{i+3} = 3P_{i+2} - 3P_{i+1} + P_i, \quad i \geq 0,$$

to generate a point sequence on a polynomial quadratic parametric curve, which is a parabola, with equally spaced parameter values. This point sequence is then mapped projectively onto a rational parametric curve, which is a general conic. By rearranging, this method can be implemented by iterative computation of forward differences, therefore it will be called the *forward differencing method* for rendering a rational quadratic curve. The advantage of this method is that the rational representation is compatible with the curve and surface representation in most current geometric modeling systems, usually including higher degree rational Bézier curves and surfaces, and rational B-spline curves and surfaces. Its shortcoming is that the point sequence generated on the conic is usually not evenly spaced. Although reparameterization or adaptively changing the step size can alleviate this problem, it will be shown later in Section 2.5 that equidistant points on a circle cannot be produced by forward differencing. Point sequences generated by forward differencing can also be generated by subdivision based on the deCasteljau algorithm [Far88], which is numerically more stable but, at the same time, less efficient than forward differencing.

The matrix iteration method is another way to generate points on a conic. It performs the iteration

$$P_{i+1} = AP_i, \quad i \geq 0, \tag{2.1}$$

for some initial point P_0 in the plane and a 2×2 real matrix A. It is shown in [Pav83] that the resulting point sequence $\{P_i\}_{i \geq 0}$ can be made to lie on any central conic with the center at the origin by choosing P_0 and A properly. Therefore, by applying a translation, any central conic can be dealt with by this method. The characteristic of the matrix A used for generating a point sequence on a central conic or a straight line is that $\det(A) = 1$.

The main criteria for judging a method for drawing conics are the efficiency for computing a point on the curve and the quality of the point sequence generated. For the forward differencing method, on average six additions and two divisions are needed for computing one point on a conic in E^2. For the matrix iteration method the computation count is four additions and four multiplications per point on a central conic; note that two additions are needed for the translation. Equidistant points on a circle can be generated by this method. In general, the matrix iteration method generates point sequences on conics of better quality than the forward differencing method, but is less efficient than the latter on computers which perform multiplications and divisions at about the same speed.

The method we shall discuss is based on the difference equation

$$P_{i+3} = (2k+1)(P_{i+2} - P_{i+1}) + P_i, \quad i \geq 0,$$

where k is an appropriate constant. It can be shown that this iteration can generate points on any proper conic; a proper conic is also called a nondegenerate conic. This method will be called the *difference method*, and k the *difference parameter*. For this method the computation count is four additions and two multiplications for generating a point on a conic. It will be shown that any point sequence generated by the matrix iteration method can also be generated by the difference method. So the difference method has the efficiency of the forward differencing method and the quality of the point sequence generated by the matrix iteration method. The difference method was first used to draw ellipses using a microprocessor controlled plotter [Wan86]. Error analysis for this method has been discussed in [WaW89a] and [WaW89b]. In this chapter, more properties and implementation issues of the difference method will be discussed.

The main features of the difference method include:

1. It is uniform for drawing all conics.
2. The computation involved in each iteration is simple; only two multiplications and four additions are required. This is the most efficient algorithm for computing a point sequence on a conic.
3. The generated polygon, as an approximation to the conic, possesses a best approximation property, which will be explained later.
4. The form of the difference equation is invariant under affine transformations.

This chapter is organized as follows. Section 2.2 studies the difference equation satisfied by point sequences on conics. In Section 2.3 we determine the difference parameter k such that the generated polygon approximates the conic within a prescribed tolerance. In Section 2.4 we consider applying the difference method to draw conics represented by the rational quadratic parameterization and the quadratic algebraic equation. In Section 2.5 we derive several properties of the difference method.

2.2 Difference method

In this section we develop the theory of the difference equation satisfied by a point sequence on a conic. We will first consider a second order difference equation in the case of central conics with center at the origin, and then consider the case of general conics. A point in the plane E^2 is represented by Cartesian coordinates $P = [x, y]^T$. The segment joining points A and B is denoted by $\overline{AB}$. When $A \neq B$ the straight line passing through A and B is denoted by AB. The vector from point A to point B is denoted by $\overrightarrow{AB} = B - A$. The length of vector $V = [x, y]^T$ is denoted by $|V| = \sqrt{V^T V}$. A continuous finite arc on a conic is called a *conic arc*, or more specifically, an *elliptic arc*, a *parabolic arc* or a *hyperbolic arc*. Ellipses and hyperbolas are *central conics*; and elliptic and hyperbolic arcs are called *central arcs*. The center of a central arc is the center of the underlying conic. The apices of a conic are the points that have the maximum curvature. We assume that a point sequence on a hyperbola lies on a branch of the hyperbola. This assumption does not pose any problem in applications and is adopted for simplicity.

2.2.1 Conics with center at origin

Now we review the trigonometric parameterizations of ellipses and hyperbolas. Suppose that C is an ellipse or a branch of a hyperbola. When C is an ellipse it can be mapped into the unit circle U by an affine transformation, where U: $x^2 + y^2 = 1$ can be expressed in the parametric form

$$P(t) = [\cos t, \sin t]^T, \quad -\infty < t < \infty. \tag{2.2}$$

Although it suffices to restrict t to any interval of length 2π, the extended parameter range will be needed later. Eqn. (2.2) will be called the *canonical form of ellipses.* When ellipse C has its center at the origin it can be expressed as

$$P(t) = R\,[\cos t,\ \sin t]^T, \quad -\infty < t < \infty, \tag{2.3}$$

where R is a 2×2 nonsingular real matrix.

When C is a branch of a hyperbola it can be brought into the hyperbolic branch H by an affine transformation, where H: $x^2 - y^2 = 1$, $x > 0$, can be expressed in the parametric form

$$P(t) = [\cosh t,\, \sinh t]^T, \quad -\infty < t < \infty. \tag{2.4}$$

Eqn. (2.4) will be called the *canonical form of hyperbolas.* When hyperbola H has center at the origin it can be expressed as

$$P(t) = R\,[\cosh t,\, \sinh t]^T, \quad -\infty < t < \infty, \tag{2.5}$$

where R is a 2×2 nonsingular real matrix.

The following theorem is useful in drawing a central conic with center at the origin.

Theorem 2.2.1: *Let C be an ellipse or a branch of a hyperbola with center at the origin O. Let P_0 and P_1 be two distinct points on C . Then there exists a constant k such that the difference equation*

$$P_{i+2} = 2kP_{i+1} - P_i, \quad i \geq 0, \tag{2.6}$$

determines a point sequence $\{P_i\}_{i \geq 0}$ on C and the P_i have equally spaced parameter values with respect to representation (2.3) or (2.5). When C is an ellipse, $-1 \leq k < 1$; when C is a hyperbola, $k > 1$. Conversely, given the initial point P_0 on an ellipse or a branch of hyperbola, and k in the range specified above for each case, there exists a point P_1 on C such that the point sequence generated by (2.6) is on C.

PROOF: Ellipses and hyperbolas will be treated separately.

Case 1: C is an ellipse. First consider the unit circle U: $x^2 + y^2 = 1$, which can be written in canonical form (2.2). Let $P_0 = [\cos t_0, \sin t_0]^T$, $P_1 = [\cos(t_0 + \theta), \sin(t_0 + \theta)]^T$ be two distinct points on U. Using the following identities

$$\cos(\alpha + \beta) + \cos(\alpha - \beta) = 2\cos\beta\cos\alpha,$$

$$\sin(\alpha + \beta) + \sin(\alpha - \beta) = 2\cos\beta\sin\alpha,$$

it is easy to show that

$$P_{i+2} = 2kP_{i+1} - P_i, \quad i \geq 0,$$

with $k = \cos\theta$, generates the point sequence $\{P_i\}_{i \geq 0}$, where $P_i = [\cos(t_0 + i\theta), \sin(t_0 + i\theta)]^T$, which obviously lies on the circle U.

Now consider ellipse C centered at the origin O. The unit circle U can be mapped onto C by a homogeneous affine transformation, denoted by a nonsingular matrix R. For any two distinct points Q_0 and Q_1 on C, choose $P_0 = R^{-1}Q_0$, $P_1 = R^{-1}Q_1$ on U. Suppose that $\{P_i\}_{i \geq 0}$ is the point sequence on U obtained from (2.6) with P_0, P_1 as initial points, and with $k = \cos\theta$ as defined above. Let $Q_i = RP_i$, $i \geq 2$. Then $\{Q_i\}_{i \geq 0}$ is a point sequence on

C satisfying (2.6). This can be verified by multiplying R to both sides of (2.6). Since P_0 and P_1 are distinct and $k = \cos\theta$, we have $-1 \leq k < 1$.

Case 2: C is one branch of a hyperbola. The proof is similar to that for the ellipse. Here one starts with the canonical form (2.4) and uses the identities

$$\cosh(\alpha+\beta) + \cosh(\alpha-\beta) = 2\cosh\beta\cosh\alpha, \tag{2.7}$$

$$\sinh(\alpha+\beta) + \sinh(\alpha-\beta) = 2\cosh\beta\sinh\alpha. \tag{2.8}$$

In this case $k = \cosh\theta > 1$ since $\theta \neq 0$.

The second part of the theorem follows easily from the above analysis. Assuming $P(t)$ is in the form (2.3) or (2.5) depending on the type of C, and $P_0 = P(t_0)$. Choose $P_1 = P(t_0+\theta)$ where θ is determined by $\cos\theta = k$ or $\cosh\theta = k$. Then the point sequence generated by (2.6) is $P_i = P(t_0 + i\theta)$, which is on C. □

Because two solutions for θ can be found from $\cos\theta = k$ (or $\cosh\theta = k$), there are two different choices of P_1, which determine the two different directions of $\{P_i\}_{i\geq 0}$ on C. We remark that in (2.6), when k is chosen such that $k < -1$, the points generated by (2.6) alternate between the two branches of a hyperbola; they are on a straight line if $k = 1$, and alternate between two parallel lines if $k = -1$. In fact, the point sequences produced by (2.6) always lie on either central conics (ellipses or hyperbolas) or straight lines. These properties can be derived easily by analyzing the characteristic equation of (2.6) [WaW89b].

2.2.2 General conics

To encompass the case of parabolas by the same difference equation as for ellipses and hyperbolas, we shall consider a difference equation of third order. In this section the centers of the conics are not necessarily at the origin O.

Let S be the center of a central conic C with parameterization $P(t)$, and let $\bar{C}$ be the conic represented by $V(t) = P(t) - S$. Then $\bar{C}$ is a translation of C and its center is at the origin O. Let $\{V_i\}_{i\geq 0}$ be a point sequence generated by (2.6) on $\bar{C}$ with an appropriately specified k, i.e.

$$V_{i+2} = 2kV_{i+1} - V_i, \quad i \geq 0.$$

Let $P_i = V_i + S$. Then a point sequence $\{P_i\}_{i\geq 0}$ on C is derived from $\{V_i\}_{i\geq 0}$. Substituting $V_i = P_i - S$ in the above equation and rearranging , we have

$$P_{i+2} = 2kP_{i+1} - P_i + 2(1-k)S, \tag{2.9}$$

and

$$P_{i+3} = 2kP_{i+2} - P_{i+1} + 2(1-k)S,$$

Eliminating $2(k-1)S$ by subtracting the above two equations, we have

$$P_{i+3} = (2k+1)P_{i+2} - (2k+1)P_{i+1} + P_i, \quad i \geq 0.$$

This is the difference equation satisfied by a point sequence $\{P_i\}_{i\geq 0}$ on C, which is a central conic with its center not necessarily at the origin.

Now we can see how the difference equation for parabolas fits into the above equation. A parabola C is expressible in the following parametric form $P(t) = [x(t), y(t)]^T$ where

$$\begin{cases} x(t) = a_1t^2 + b_1t + c_1 \\ y(t) = a_2t^2 + b_2t + c_2, \end{cases} \quad -\infty < t < \infty. \tag{2.10}$$

Suppose that on C there is a point sequence $\{P_i\}_{i\geq 0}$ given by $P_i = P(t_0 + i\theta)$, for some t_0 and $\theta \neq 0$. As the third order forward difference of any quadratic polynomial is identically zero, we have

$$P_{i+3} - 3P_{i+2} + 3P_{i+1} - P_i = 0.$$

Hence

$$P_{i+3} = 3P_{i+2} - 3P_{i+1} + P_i, \quad i \geq 0. \tag{2.11}$$

This is essentially the same equation as used in the forward differencing method. By an appropriate arrangement, the arithmetic operations required by the forward differencing method are two additions per coordinate for a new point [BBB87].

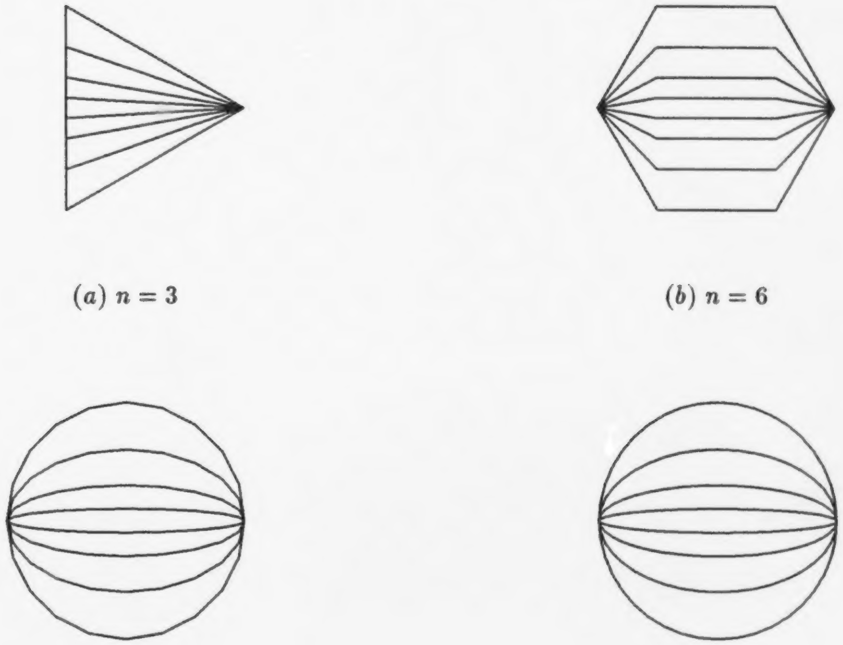
We summarize the above results in the following theorem.

Theorem 2.2.2: *On a conic C, for any initial point P_0 and constant k which is in a range depending on the type of C as indicated below, there exist points P_1 and P_2 such that*

$$P_{i+3} = (2k+1)P_{i+2} - (2k+1)P_{i+1} + P_i, \quad i \geq 0, \tag{2.12}$$

generates a point sequence $\{P_i\}_{i\geq 0}$ that lies on C. For ellipses $-1 \leq k < 1$; for parabolas $k = 1$; for hyperbolas $k > 1$. For a parabola the distribution and direction of the sequence $\{P_i\}_{i\geq 0}$ on C are solely determined by P_1. For ellipses and hyperbolas the distribution is determined by k, and the direction by P_1.

Later in this chapter we will see how the results of this theorem are applied to drawing conics. Fig. 2.1.1 illustrates some polygonal approximations generated by the difference method for four ellipses. An advantage of (2.12) over (2.6) is its invariance under any affine transformation, while (2.6) is invariant only under homogeneous affine transformations.



(*a*) $n = 3$ (*b*) $n = 6$

(*c*) $n = 20$ (*d*) $n = 40$

Figure 2.2.1 Four ellipses: Polygonal approximations of four ellipses generated by the difference method. n is the number of sides in each approximation.

By (2.12),

$$P_{i+3} - 2kP_{i+2} + P_{i+1} = P_{i+2} - 2kP_{i+1} + P_i, \quad i \geq 0.$$

Therefore

$$P_{i+2} - 2kP_{i+1} + P_i = c, \quad i \geq 0,$$

with c being a constant. This is another formula for generating a conic. When C is a central conic, we have (2.9), i.e.

$$c = 2(1-k)S, \quad i \geq 0.$$

This equation should be interpreted appropriately when the conic under consideration is a parabola, for which S is regarded as being at infinity and $1 - k = 0$. A consequence of this relation is that when S is at the origin O, (2.12) reduces to (2.6).

2.2.3 Equivalence with matrix iteration method

The matrix iteration method for drawing conics has been mentioned in Section 2.1. A detailed discussion of this method can be found in [Pav83]. The characteristic of the matrix iteration method is that the determinant of the matrix equals 1, in order to produce a successive point sequence on a central conic or a straight line. The relationship between the difference method and the matrix iteration method is explained by the following theorem.

Theorem 2.2.3: *Suppose that $\{P_i\}_{i\geq 0}$ is a point sequence on a proper conic centered at the origin but not contained in a straight line. Then $\{P_i\}_{i\geq 0}$ is generated by the matrix iteration method if and only if it is generated by the difference method.*

PROOF: Suppose that $\{P_i\}_{i\geq 0}$ is generated by the matrix method, that is, there exists a 2×2 real matrix A with $\det(A) = 1$, such that

$$P_{i+1} = AP_i, \; i \geq 0.$$

Let the characteristic polynomial of A be

$$f(\lambda) = \det(\lambda I - A) = \lambda^2 - \mathrm{tr}(A)\lambda + \det(A) = \lambda^2 - \mathrm{tr}(A)\lambda + 1.$$

By the Cayley-Hamilton Theorem,

$$f(A) = A^2 - tr(A)A + I = 0.$$

So

$$f(A)P_i = A^2P_i - \mathrm{tr}(A)AP_i + P_i = P_{i+2} - \mathrm{tr}(A)P_{i+1} + P_i = 0,$$

that is,

$$P_{i+2} = \mathrm{tr}(A)P_{i+1} - P_i, \quad i \geq 0.$$

Let $\mathrm{tr}(A) = 2k$. Thus $\{P_i\}_{i\geq 0}$ can be produced by the difference method.

Conversely, let $\{P_i\}_{i\geq 0}$ be a point sequence generated by the difference method (2.6). Then there exists constant k such that

$$P_{i+2} = 2kP_{i+1} - P_i, \quad i \geq 0.$$

$$[P_1\ P_2] = [P_0\ P_1] \begin{bmatrix} 0 & -1 \\ 1 & 2k \end{bmatrix}. \tag{2.13}$$

By the given condition on $\{P_i\}_{i\geq 0}$, we can assume that P_0 and P_1 are linearly independent. Let

$$A = [P_0\ P_1] \begin{bmatrix} 0 & -1 \\ 1 & 2k \end{bmatrix} [P_0\ P_1]^{-1}.$$

Then

$$A[P_0\ P_1] = [P_1\ P_2]. \tag{2.14}$$

Now the proof is to be completed by induction. Eqn. (2.14) gives the base cases for $i = 0$ and 1. Suppose for any $i \leq l$, we have $P_{i+1} = AP_i$, then for $i = l + 1$

$$P_{l+2} = 2kP_{l+1} - P_l = 2kAP_l - AP_{l-1} = A(2kP_l - P_{l-1}) = AP_{l+1}.$$

By induction we have proved that $P_{i+1} = AP_i$ for any $i \geq 0$, i.e. $\{P_i\}_{i\geq 0}$ can be generated by the matrix method. □

2.2.4 Approximation property and efficiency

The inscribed polygon of a conic arc generated by the matrix iteration method is discovered in [Smi71] to possess the following best approximation property of maximum inscribed area. Let P_i, $i = 0, 1, \ldots, n$, be the point sequence generated by the difference method on a conic arc C with P_0 and P_n being the two endpoints of C. Then connecting P_0 and P_n gives a convex polygon $G : P_0P_2 \ldots P_n$. The best approximation property of maximum area states that the polygon G has the maximum area among all the convex $n+1$ sided polygons whose vertices are on the arc C and include the two endpoints of C. In [Smi71] Smith uses the matrix iteration method for axis-aligned conics only, but an affine transformation does not change the property of maximum area, so by Theorem 2.2.3, the inscribed polygons of the conic arcs generated by (2.6), as well as (2.12), also possess the same best approximation property as those generated by the matrix iteration method. However, the proof given in

[Smi71] for the above property contains an incorrect argument. We shall give a proof of this property in Section 2.5.

The efficiency of the difference method in the plane E^2 has been mentioned earlier. The saving of operation counts by this method compares more favorably with the matrix iteration method in higher dimensional space. Because the difference method represented by (2.12) is invariant under affine transformations, the difference equation of a point sequence on a conic in E^d has the same form (2.12), with each point having d components. In E^d, $d \geq 2$, the difference method needs d multiplications and $2d$ additions/subtractions for computing one point. The higher dimensional version of the matrix iteration method, which generates the same point sequences as the difference method, is obtained by replacing A in (2.1) by a $d \times d$ matrix. It needs d^2 multiplications and d^2 additions, including translation, to generate a point.

Finally, an observation of importance to applications is that if the value of the difference parameter k is assumed to be of the form 1 ± 2^{-s}, where s is a positive integer, then the computation of each point using (2.6) or (2.12) can be carried out with only additions and shifts. This means that the inscribed polygon of any conic can be computed without using multiplications. This feature of the difference method enables it to be implemented in VLSI easily to speed up the rendering of conics drastically. This advantage could also be utilized when the algorithm for drawing conic arcs is coded in machine code or assembly language, in which shifts can be done efficiently.

2.3 Drawing conic arcs to desired tolerance

In this section we will discuss one of the implementation issues of the difference method. The problem is: given the equation of a conic arc in some form, two endpoints of the arc, and a tolerance of approximation μ, determine a difference parameter k so that the distances from the sides of the generated polygon to the arc are not greater than μ.

The determination of the difference parameter k is fundamental to the application of the difference method. By Theorem 2.2.2, k affects the distribution (or density) of the point sequence generated by the corresponding difference equation. So when a precision requirement is imposed, we need to find out a suitable difference parameter meeting the requirement. Intuitively, if the distribution of the point sequence on a curve is dense enough, the approximation should be good. But a high density of points generated on the arc could cause excessive unnecessary computation.

We shall discuss how to specify the precision requirement of approximation, and how to determine an appropriate difference parameter to satisfy the requirement. The main result is a simple formula which determines an appropriate difference parameter, once some simple geometric characteristics of the conic arc are known and the precision requirement is given. Since $k = 1$ for all parabolas, we shall address parabolas separately.

The following theorem is the preparation for our analysis.

Theorem 2.3.1: *Suppose that C is a proper elliptic or hyperbolic arc with center at the origin O, and $\{P_i\}_{i=0}^n$ is a point sequence on C generated by the difference method. On the arc P_iP_{i+1}, let $P_{i+1/2}$ be the point with the maximum distance to the segment $\overline{P_iP_{i+1}}$, and let this maximum distance μ_i be called the* chord-arc offset *on the segment $\overline{P_iP_{i+1}}$. Then (1) when C is elliptic: $\mu_i > \mu_j$ iff $|P_{i+1/2}| > |P_{j+1/2}|$; (2) when C is hyperbolic: $\mu_i > \mu_j$ iff $|P_{i+1/2}| < |P_{j+1/2}|$.*

In the above $|P|$ denotes the distance of point P to the origin O which has been assumed to be the center of C. Theorem 2.3.1 can be used to determine the location where the maximum chord-arc offset on a central arc C occurs. For instance, the theorem implies that when C is elliptic, if the middle point $P_{i+1/2}$ of arc P_iP_{i+1} is farthest to the center among all the $|P_{j+1/2}|$, $j = 0, 1, \ldots, n-1$, then the corresponding chord-arc offset is the largest.

To prove Theorem 2.3.1, we first need the following lemmas.

Lemma 2.3.2: *Let $\{P_i\}_{i=0}^n$ be a point sequence generated by the difference equation (2.12) on a central conic C. Then the area enclosed by the arc P_iP_{i+1} and the segment $\overline{P_iP_{i+1}}$ is the same for all $i \geq 0$.*

PROOF: Since the ratio of areas is invariant under affine transformations, we just have to prove the lemma for central conics in canonical form.

Case 1: C is elliptic. Although in this case the statement is obvious from a geometric point of view, we will give a computational proof since the result of the computation will be useful later in this chapter.

The canonical form of an ellipse is given by (2.2). As indicated in the proof of Theorem 2.2.1, the point sequence $\{P_i\}_{i\geq 0}$ generated by (2.12), or equivalently by (2.6) since C has its center at the origin, is

$$P_i = [\cos(t_0 + i\theta), \sin(t_0 + i\theta)]^T.$$

Let $S(\overline{P_iP_{i+1}})$ be the area enclosed by the arc P_iP_{i+1} and the segment $\overline{P_iP_{i+1}}$. Let the parameterization of the segment $\overline{P_iP_{i+1}}$ be $B(s) = [x(s), y(s)]^T = sP_i + (1-s)P_{i+1}$, where $0 \le s \le 1$. By Green's Theorem [O'Neil, p. 261], the area $S(\overline{P_iP_{i+1}})$ can be calculated by the following formula.

$$\begin{aligned} S(\overline{P_iP_{i+1}}) &= \frac{1}{2}\left|\int_{t_0+i\theta}^{t_0+(i+1)\theta}[x(t)dy(t) - y(t)dx(t)] + \int_0^1 [x(s)dy(s) - y(s)dx(s)]\right| \\ &= \frac{1}{2}\left|\int_{t_0+i\theta}^{t_0+(i+1)\theta}(\sin^2 t + \cos^2 t)dt\right. \\ &\quad \left. + \int_0^1 [\cos(t_0+(i+1)\theta)\sin(t_0+i\theta) - \cos(t_0+i\theta)\sin(t_0+(i+1)\theta)]ds\right| \\ &= \frac{1}{2}|\theta - \sin\theta|, \quad i \ge 0. \end{aligned}$$

Since $S(\overline{P_iP_{i+1}})$ is independent of i, the lemma has been proved for ellipses in canonical form.

Case 2: C is hyperbolic. The canonical form for hyperbolas is given by (2.4). The point sequence generated by (2.12) is

$$P_i = [\cosh(t_0 + i\theta), \sinh(t_0 + i\theta)]^T, \quad i \ge 0.$$

Through a calculation similar to that in *case* 1, we obtain

$$S(\overline{P_iP_{i+1}}) = \frac{1}{2}|\sinh\theta - \theta|, \quad i \ge 0,$$

which is independent of i. So the lemma is true for hyperbolas in canonical form. □

A similar argument shows that the above lemma is also true for a point sequence generated by the difference method on a parabola.

Lemma 2.3.3: *Let* $\{P_i\}_{i\ge 0}$ *be a point sequence generated on a central conic* C *with center at the origin by the difference equation* $P_{i+2} = 2kP_{i+1} - P_i$, $i \ge 0$. *Then there exists* $\{Q_i\}_{i\ge 0}$ *on* C *satisfying*

$$Q_{i+2} = 2k'Q_{i+1} - Q_i, \quad i \ge 0, \tag{2.15}$$

with $k' = [(k+1)/2]^{1/2}$ *such that* $Q_{2i} = P_i$, $i \ge 0$.

PROOF: By Theorem 2.2.1, let $P_i = P(t_0 + i\theta)$, $i \ge 0$, for some t_0 and $\theta \ne 0$, where $P(t)$ is the parametric form (2.3) or (2.5) of the arc C. Define $Q_i = P(t_0 + i\theta/2)$, $i \ge 0$. Then it can be verified that the Q_i's satisfy (2.15) with $k' = [(k+1)/2]^{1/2}$, for in the case of ellipses

$$k' = \cos(\theta/2) = [(\cos\theta + 1)/2]^{1/2} = [(k+1)/2]^{1/2};$$

and in the case of hyperbolas

$$k' = \cosh(\theta/2) = [(\cosh\theta + 1)/2]^{1/2} = [(k+1)/2]^{1/2}.$$

Obviously, $Q_{2i} = P_i$, $i \geq 0$. □

Lemma 2.3.4: *If* $\{P_i\}_{i\geq 0}$ *is a point sequence generated by the difference method on a proper conic, then* P_{i+1} *is the unique point on the arc* P_iP_{i+2} *that yields the maximum distance from the arc* P_iP_{i+2} *to line segment* $\overline{P_iP_{i+2}}$.

PROOF: When the point sequence is generated by the matrix iteration method on a proper central conic, the conclusion of this lemma is proved in [Smi71]. From Theorem 2.2.3 we know that on a central conic the difference method generates the same point sequence as the matrix iteration method. So the lemma is true for a central conic. For a parabolic arc the lemma is proved directly in [Smi71]. □

PROOF of Theorem 2.3.1: By Lemma 2.3.3, there exists a point sequence $\{Q_i\}_{i\geq 0}$ generated by the difference method with appropriate initial points and parameter k' such that $Q_{2i} = P_i$. By Lemma 2.3.4, Q_{2i+1} is the unique point on the arc P_iP_{i+1} that yields the maximum distance to segment $\overline{P_iP_{i+1}}$, therefore $Q_{2i+1} = P_{i+1/2}$ by the definition of $P_{i+1/2}$.

Let the area of triangle $\Delta P_iQ_{2i+1}P_{i+1}$ be $S(\Delta P_iQ_{2i+1}P_{i+1})$, then

$$S(\Delta P_iQ_{2i+1}P_{i+1}) = S(\overline{P_iP_{i+1}}) - S(\overline{P_iQ_{2i+1}}) - S(\overline{Q_{2i+1}P_{i+1}}) \equiv Area, \tag{2.16}$$

recalling that $S(\overline{P_iP_{i+1}})$ is the area enclosed by the arc P_iP_{i+1} and segment $\overline{P_iP_{i+1}}$. By Lemma 2.3.2, $Area$ is independent of i. Let $d_i = |\overline{P_iP_{i+1}}|$. Then what was said above is equivalent to $\frac{1}{2}\mu_i d_i = Area$, where μ_i is the distance from $P_{i+1/2}$ to $\overline{P_iP_{i+1}}$. Thus $\mu_i = 2Area/d_i$. It follows that $\mu_i > \mu_j$ iff $d_i < d_j$. So in order to prove the theorem it suffices to show that (1) $d_i < d_j$ iff $|P_{i+1/2}| > |P_{j+1/2}|$ when C is elliptic; (2) $d_i < d_j$ iff $|P_{i+1/2}| < |P_{j+1/2}|$ when C is hyperbolic.

Since the Euclidean distance d_i is invariant under orthogonal transformations, for convenience, we just need to prove the theorem assuming C in the form

$$P(t) = [a\cos t, b\sin t]^T, \quad -\infty < t < \infty$$

or

$$P(t) = [a\cosh t, b\sinh t]^T, \quad -\infty < t < \infty,$$

for some a, $b > 0$, depending on whether C is elliptic or hyperbolic.

Case 1: C is elliptic. The solution of $\{P_i\}_{i\geq 0}$, given by (2.6), is

$$P_i = [a\cos(t_0 + i\theta), b\sin(t_0 + i\theta)]^T, \quad i \geq 0.$$

So

$$d_i^2 = |P_iP_{i+1}|^2$$

$$\begin{aligned}
&= [(a\cos(t_0+(i+1)\theta)-a\cos(t_0+i\theta)]^2+[(b\sin(t_0+(i+1)\theta)-b\sin(t_0+i\theta)]^2\\
&= 4a^2\sin^2\frac{\theta}{2}\sin^2(t_0+\frac{2i+1}{2}\theta)+4b^2\sin^2\frac{\theta}{2}\cos^2(t_0+\frac{2i+1}{2}\theta).\\
&= 4\sin^2\frac{\theta}{2}[a^2+b^2-(a^2\cos^2(t_0+\frac{2i+1}{2}\theta)+b^2\sin^2(t_0+\frac{2i+1}{2}\theta))].
\end{aligned}$$

On the other hand, since $\{Q_i\}_{i\geq 0}$ satisfies

$$Q_{i+2}=2k'Q_{i+1}-Q_i,\quad i\geq 0,$$

where $k'=\cos\theta/2$ and $Q_{2i}=P_i$, we have

$$Q_i=[a\cos(t_0+\frac{i\theta}{2}),b\sin(t_0+\frac{i\theta}{2})]^T.$$

Therefore

$$d_i^2=4\sin^2(\theta/2)[a^2+b^2-|Q_{2i+1}|^2]=4\sin^2(\theta/2)[a^2+b^2-|P_{i+1/2}|^2],\quad i\geq 0.$$

This equality implies that $d_i<d_j$ iff $|P_{i+1/2}|>|P_{j+1/2}|$.

Case 2: C is hyperbolic. Since in this case

$$P_i=[a\cosh(t_0+i\theta),b\sinh(t_0+i\theta)]^T,\quad i\geq 0,$$

and

$$Q_i=[a\cosh(t_0+\frac{i\theta}{2}),b\sinh(t_0+\frac{i\theta}{2})]^T,\quad i\geq 0,$$

we have

$$\begin{aligned}
d_i^2 &= 4\sinh^2(\theta/2)[b^2-a^2+a^2\cosh^2(t_0+\frac{2i+1}{2}\theta)+b^2\sinh^2(t_0+\frac{2i+1}{2}\theta)]\\
&= 4\sinh^2(\theta/2)(b^2-a^2+|Q_{2i+1}|^2)\\
&= 4\sinh^2(\theta/2)(b^2-a^2+|P_{i+1/2}|^2),\quad i\geq 0.
\end{aligned}$$

This equality implies that $d_i<d_j$ iff $|P_{i+1/2}|<|P_{j+1/2}|$. □

Now we are able to derive a formula for determining the difference parameter k for a point sequence $\{P_i\}_{i=0}^n$, given a tolerance μ as an upper bound on the chord-arc offsets resulting from $\{P_i\}_{i=0}^n$. Only the case of hyperbolas will be analyzed in detail; the case of ellipses follows from similar arguments.

Let $\{P_i\}_{i=0}^n$ be a point sequence generated by the difference method with difference parameter k on a hyperbolic arc C with center at the origin O. Then, by Lemma 2.3.3, there exists sequence $\{Q_i\}_{i=0}^{2n}$ satisfying the difference equation (2.15) with difference parameter $k'=[(k+1)/2)]^{1/2}$ such that $Q_{2i}=P_i$. Let μ be the tolerance imposed as an upper

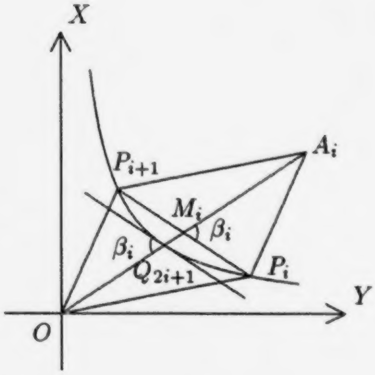


Figure 2.3.1 A hyperbolic arc.

bound on the chord-arc offsets μ_i resulting from $\{P_i\}_{i\geq 0}$. Let $M_i = (Q_{2i} + Q_{2i+2})/2$ be the midpoint of $\overline{Q_{2i}Q_{2i+2}}$. See Fig. 2.3.1 for an illustration. From (2.15) we have

$$M_i - Q_{2i+1} = \frac{1}{2}(Q_{2i} + Q_{2i+2}) - \frac{1}{2k'}(Q_{2i} + Q_{2i+2}) = (k' - 1)Q_{2i+1}.$$

Let $\beta_i = \angle OM_iQ_{2i}$, or equivalently, by Lemma 2.3.4, the angle formed by the line segment $\overline{OQ}_{2i+1}$ and the tangent of C at Q_{2i+1}. For our purpose the two such mutually supplementary angles have the same effect, since we will only be concerned with $\sin\beta_i$. According to Lemma 2.3.2 the offset of $\overline{P_iP_{i+1}}$ from the arc P_iP_{i+1} is attained at Q_{2i+1} and is $\mu_i = |M_i - Q_{2i+1}|\sin\beta_i$. If it is demanded that the offset on the segment $\overline{P_iP_{i+1}}$ does not exceed the prescribed tolerance μ, then

$$\begin{aligned}\max_i\{\mu_i\} &= \max_i\{|M_i - Q_{2i+1}|\sin\beta_i\} \\ &= \max_i\{(k'-1)|Q_{2i+1}|\sin\beta_i\} \leq \mu. \qquad (2.17)\end{aligned}$$

By Theorem 2.3.1, $\max\{\mu_i\}$ occurs at the Q_{2i+1} with the smallest $|Q_{2i+1}|$. Let E_0 and E_1 be endpoints of the arc C, let A be the apex of the underlying hyperbolic branch of C, and let $\beta(X)$ be the angle formed by the line segment $\overline{OX}$ with the tangent of C at X, where $X = E_0, E_1$, or A. Then we have

$$\max_i\{\mu_i\} \leq (k'-1)\max\{|E_0|\sin\beta(E_0), |E_1|\sin\beta(E_1)\} \qquad (2.18)$$

if A is not on C, since in this case E_0 or E_1 is closest to the center among all points on arc C; and

$$\max_i\{\mu_i\} \leq (k' - 1)|A| \sin\beta(A) \tag{2.19}$$

if A is on C, since in this case A is closest to the center among all points on arc C.

To find k' from (2.17), it suffices to set the right hand side of (2.18) or (2.19) $\leq \mu$. Thus k' can be chosen to be

$$k' = 1 + \mu/L(C), \tag{2.20}$$

where $L(C) = \max\{|E_0|\sin\beta(E_0), |E_1|\sin\beta(E_1)\}$ if the apex A is not on C; or $L(C) = |A|\sin\beta(A)$ if A is on C. Finally the difference parameter k of the sequence $\{P_i\}_{i\geq 0}$, by Lemma 2.3.3, is

$$k = 2k'^2 - 1 \tag{2.21}$$

where k' is given by (2.20).

For ellipses the reasoning is similar to that for hyperbolas. So we will only give the formula without proof. For ellipses the counterpart of (2.20) becomes

$$k' = 1 - \mu/L(C), \tag{2.22}$$

where $L(C)$ is defined as above with A being one of the apices of the underlying ellipse of the arc C. With k' being known, k is also given by (2.21).

Summarizing the above results, we see that if the difference parameter k is determined as above, then the generated inscribed polygon of C approximates C within the prescribed tolerance μ. Since the choice of k depends only on the geometric properties of C, the above results also apply to conic arcs with center not at the origin.

It is easy to see that the quantity $|X|\sin\beta(X)$ appearing in the above equations is the distance $d(X)$ from the center of the conic section to the tangent of C at X, where $X = E_0$, E_1, or A. This fact further simplifies the computation of k, since $d(X)$ can now be calculated easily from either the parametric or the implicit equation of a conic arc. This observation about $|X|\sin\beta(X)$ also helps to explain the conclusion of Theorem 2.3.1.

Now we need to do the above analysis for parabolas. Let the parametric equation of the parabola be

$$P(t) = Ut^2 + Vt + W. \tag{2.23}$$

We assume that $\{P(t_i)\}_{i\geq 0}$, where $t_i = t_0 + i\theta$, is the point sequence produced by (2.12) with the difference parameter $k = 1$. Since the difference parameter $k = 1$ for all parabolas, now we need to find the relation between the parameter increment θ for successive points

$\{P_i\}_{i\geq 0}$ and the degree of approximation. By Lemma 2.3.4, the offset μ_i of the arc P_iP_{i+1} to the segment $\overline{P_iP_{i+1}}$ is attained at $P_{i+1/2} = P(t_i + \theta/2)$. So

$$\begin{aligned}
\mu_i &= \frac{|[P(t_i+\theta/2) - P(t_i)] \times (P(t_{i+1}) - P(t_i))|}{|P(t_{i+1}) - P(t_i)|} \\
&= \frac{|[(t_i\theta + \theta^2/4)U + \theta V/2] \times [(2t_i\theta + \theta^2)U + \theta V)|}{|(2t_i\theta+\theta^2)U + \theta V|} \\
&= \frac{1}{4}\frac{\theta^2|U \times V|}{|2(t_i+\theta/2)U + V|} \\
&= \frac{\theta^2}{4}\frac{|U\times V|}{|P'(t_i+\theta/2)|}. \qquad (2.24)
\end{aligned}$$

On the other hand the curvature of the parabola at $P(t)$ is

$$\kappa(t) = \frac{|P'(t)\times P''(t)|}{|P'(t)|^3} = \frac{|(2tU+V)\times(2U)|}{|P'(t)|^3} = \frac{2|U\times V|}{|P'(t)|^3}.$$

Thus, replacing t by $t+\theta/2$,

$$|P'(t+\theta/2)| = \left[\frac{2|U\times V|}{\kappa(t+\theta/2)}\right]^{1/3}.$$

Substituting this in the above expression for μ_i, we have

$$\mu_i = \frac{\theta^2}{4}\left[\frac{|U\times V|^2\kappa(t_i+\theta/2)}{2}\right]^{1/3}.$$

Since on a parabola the maximum curvature is attained at the apex, μ_i attains its maximum over a parabolic arc C at the endpoints of the arc if the apex is not contained on the arc C, or at the apex if it is on the arc C. This result is comparable with the similar property of ellipses and hyperbolas stated in Theorem 2.3.1.

When a bound μ on the μ_i's is prescribed, by (2.24) the parameter increment θ is determined by

$$|\theta| = 2\sqrt{\mu L(C)/|U\times V|}, \qquad (2.25)$$

where $L(C) = \min\{|E_0'|, |E_1'|\}$ when the apex A of the parabola is not on C; $L(C) = |A'|$ when A is on C. The notation X' denotes the tangent vector of the parabola at point X with respect to its parametric equation (2.23). By (2.25), the choice of θ depends on the particular parameterization $P(t)$.

2.4 Conversion from parametric and implicit forms of conics

In applications usually a conic arc is given in rational quadratic parametric form $P(t)$ or implicit form $f(x, y) = 0$, with two endpoints specified. In order to apply the difference method, we must be able to efficiently derive all necessary parameters, such as the parameter range of the arc with respect to its canonical form, the position of apices of the underlying conic relative to the arc, etc. from the given information.

In this section we first consider the case where the conic arc is given in the rational quadratic parametric form. Our initial method is straightforward, but does not allow direct processing of the semi-ellipse. Then we introduce a new parameterization for elliptic arcs which encompasses the semi-ellipse. Finally we deal with the case where the conic arc is given in the implicit form $f(x, y) = 0$.

2.4.1 Conics as rational quadratic curves

The rational quadratic parametric representation of conics provides a convenient way to specify a conic arc [Far89, Lee87]. It can also be used to produce the approximate polygon of the arc using forward differencing. Since the difference method introduced in the preceding sections is more efficient and produces point sequences of better quality than forward differencing, we intend to apply the difference method to a conic arc in the rational quadratic form. To this end, we shall first address the problem of converting the rational parametric representation to a form in which it is easier to apply the difference method.

First we introduce a parametric representation with a local parameter for elliptic and hyperbolic arcs with center at the origin. Both representations are closely related to the canonical forms (2.2) or (2.4) of the underlying conic of the arc, so it is straightforward to apply the difference method to a conic arc in the new representation. Then we consider how to convert the rational representation to these new parametric representations for elliptic and hyperbolic arcs, respectively.

Theorem 2.4.1: *Let C be an elliptic or hyperbolic arc with center at the origin, and with endpoints $V_s = P(t_0)$ and $V_e = P(t_0 + \alpha)$, $\alpha > 0$, where $P(t)$ is given by (2.3) or (2.5). Then C can be expressed as*

$$V(u) = \frac{\sin[(1-u)\alpha]}{\sin\alpha} V_s + \frac{\sin(u\alpha)}{\sin\alpha} V_e, \quad 0 \le u \le 1, \tag{2.26}$$

when C is elliptic and $\sin\alpha \neq 0$*; or*

$$V(u) = \frac{\sinh[(1-u)\alpha]}{\sinh\alpha}V_s + \frac{\sinh(u\alpha)}{\sinh\alpha}V_e, \quad 0 \leq u \leq 1, \tag{2.27}$$

when C is hyperbolic.

PROOF: Consider (2.26) first. The underlying ellipse of the elliptic arc is equivalent to the circle canonical form (2.2) under a homogeneous affine transformation. Since the form of (2.26) is invariant under homogeneous affine transformations, it suffices to prove the theorem assuming that C is in canonical form, i.e. $P(t) = [x(t), y(t)]^T$ where

$$\begin{cases} x(t) = \cos t \\ y(t) = \sin t, \end{cases} \quad t_0 \leq t \leq t_0 + \alpha, \quad \alpha > 0,$$

with $V_s = P(t_0)$, and $V_e = P(t_0 + \alpha)$.

Now we must prove that the $V(u)$ defined by (2.26) is a point on arc C. We consider first the coordinate component $x(u)$ of $V(u) = [x(u), y(u)]^T$. Since $V_s = [x_s, y_s]^T = P(t_0) = [\cos t_0, \sin t_0]^T$ and $V_e = [x_e, y_e]^T = P(t_0 + \alpha) = [\cos(t_0 + \alpha), \sin(t_0 + \alpha)]^T$, substituting $x_s = \cos t_0$ and $x_e = \cos(t_0 + \alpha)$ in the right hand side of (2.26) yields

$$\begin{aligned} x(u) &= \frac{\sin[(1-u)\alpha]}{\sin\alpha}(\cos t_0) + \frac{\sin(u\alpha)}{\sin\alpha}(\cos(t_0+\alpha)) \\ &= \frac{1}{2\sin\alpha}[\sin(t_0 + (1-u)\alpha) - \sin(t_0 - (1-u)\alpha) \\ &\quad + \sin(t_0 + \alpha + u\alpha) - \sin(t_0 + \alpha - u\alpha)] \\ &= \frac{1}{2\sin\alpha}[-\sin(t_0 - (1-u)\alpha) + \sin(t_0 + \alpha + u\alpha)] \\ &= \cos(t_0 + u\alpha). \end{aligned}$$

Similarly it can be shown that the second component $y(u) = \sin(t_0 + u\alpha)$. Hence $V(u) = P(t_0 + u\alpha)$, i.e. $V(u)$ is a local parameterization of the arc C.

The proof for (2.27) follows from similar hyperbolic identities. □

The parameterization (2.26) is used in [Sho85] to represent a circular arc on a sphere.

We now consider converting a given rational quadratic representation of a conic arc to a form for which the difference method can be used to draw the arc. We will assume that the conic arc is given in rational quadratic Bézier form. Given such a conic arc, we will convert it directly into the representation (2.26) for an elliptic arc, or to (2.27) for a hyperbolic arc; the parabola can be drawn with the difference method directly since its rational representation would reduce to a quadratic polynomial form (2.10) through a rational linear reparameterization.

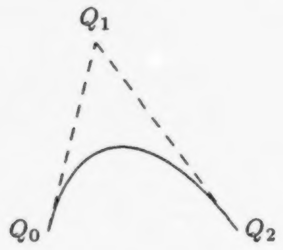


Figure 2.4.1 A rational quadratic Bézier curve with $\kappa = 25/16$.

Let C be the conic $P(t)$ in Bézier form given by

$$P(t) = \frac{w_0 B_{0,2}(t) Q_0 + w_1 B_{1,2}(t) Q_1 + w_2 B_{2,2}(t) Q_2}{w_0 B_{0,2}(t) + w_1 B_{1,2}(t) + w_2 B_{2,2}(t)}, \quad 0 \leq t \leq 1, \tag{2.28}$$

where $B_{0,2}(t) = (1-t)^2$, $B_{1,2}(t) = 2t(1-t)$, and $B_{2,2}(t) = t^2$; Q_0 , Q_1 and Q_2 are control points of the arc; w_0, w_1 and w_2 are the weights associated with the Q_i. Here we assume $w_i > 0$, i=0, 1, 2, that is, we just consider the case where the Bézier curve is contained in the control triangle $\Delta Q_0 Q_1 Q_2$. Changing the sign of w_1 gives the complementary arc of C on the underlying conic.

It is well known that $\kappa = w_0 w_2 / w_1^2$ is an invariant under both the affine coordinate transformation and the projective parameter transformation provided that the interval $[0,1]$ is mapped to $[0,1]$ and the two ends $t = 0, 1$ are fixed [Pat86]. The geometric meaning of κ is as follows. (1) $\kappa = 1$ stands for a parabolic arc; (2) $\kappa > 1$ stands for an elliptic arc; (3) $\kappa < 1$ stands for a hyperbolic arc. The center of the arc in the case $\kappa \neq 1$ is given by

$$S = Q_1 + \frac{\kappa}{2(\kappa - 1)}(Q_0 + Q_2 - 2Q_1), \tag{2.29}$$

All the above results can be found in [Lee87]. An elliptic arc in the Bézier form is shown in Fig. 2.4.1.

The following theorem is the key to converting a rational curve in the form (2.28) to the form (2.26) or (2.27).

Theorem 2.4.2: *(1) When $\kappa > 1$, the arc C given by (2.28) can be expressed as*

$$Q(u) = (Q_0 - S)\frac{\sin[(1-u)\alpha]}{\sin \alpha} + (Q_2 - S)\frac{\sin(u\alpha)}{\sin \alpha} + S, \quad 0 \leq u \leq 1, \tag{2.30}$$

where $\alpha = \arccos(\frac{2}{\kappa} - 1)$ *and* $\sin \alpha = 2\sqrt{\kappa - 1}/\kappa$.

(2) When $\kappa < 1$, the arc C given by (2.28) can be expressed as

$$Q(u) = (Q_0 - S)\frac{\sinh[(1-u)\alpha]}{\sinh\alpha} + (Q_2 - S)\frac{\sinh(u\alpha)}{\sinh\alpha} + S, \quad 0 \le u \le 1, \tag{2.31}$$

where $\alpha = \operatorname{arccosh}(\frac{2}{\kappa} - 1)$ *and* $\sinh\alpha = 2\sqrt{1-\kappa}/\kappa$.

Proof: By Theorem 2.4.1, any curve in the form (2.28) with $\kappa \neq 1$ can be represented in the form (2.30) or (2.31). Therefore we just need to determine α in terms of κ.

(1) When $\kappa > 1$, C is an elliptic arc. α is determined as follows. The derivative of $Q(u)$ in (2.30) is

$$Q'(u) = -\alpha(Q_0 - S)\frac{\cos[(1-u)\alpha]}{\sin\alpha} + \alpha(Q_2 - S)\frac{\cos(u\alpha)}{\sin\alpha},$$

and Q_1 is the intersection of the tangent lines of C at $Q_0 = Q(0)$ and $Q_2 = Q(1)$, i.e. Q_1 is the intersection of

$$T_0(s) = Q_0 + sQ'(0), \quad s \ge 0$$

and

$$T_1(v) = Q_2 - vQ'(1), \quad v \ge 0.$$

See Fig. 2.4.1. Setting $Q_1 = T_0(s) = T_1(v)$ yields

$$Q_2 - Q_0 = sQ'(0) + vQ'(1).$$

Taking the cross product with $Q'(1)$ on both sides yields

$$\begin{aligned} s &= \frac{|[(Q_2 - S) - (Q_0 - S)] \times Q'(1)|}{|Q'(0) \times Q'(1)|} \\ &= \frac{|(Q_0 - S) \times (Q_2 - S) - (Q_0 - S) \times (Q_2 - S)\cos\alpha|}{(\alpha/\sin\alpha)| - (Q_0 - S) \times (Q_2 - S)\cos^2\alpha + (Q_0 - S) \times (Q_2 - S)|} \\ &= \frac{1 - \cos\alpha}{\alpha\sin\alpha}. \end{aligned}$$

So

$$Q_1 = Q_0 + \frac{1-\cos\alpha}{\alpha\sin\alpha}Q'(0) = Q_0 + \frac{1}{1+\cos\alpha}[(Q_2 - S) - (Q_0 - S)\cos\alpha].$$

Rearranging the above equation yields

$$S = Q_1 + \frac{1}{1-\cos\alpha}[(Q_0 - Q_1) + (Q_2 - Q_1)].$$

Since $Q_0 - Q_1$ and $Q_2 - Q_1$ are linearly independent, comparing the above expression for S with (2.29), we have

$$\frac{1}{1-\cos\alpha} = \frac{\kappa}{2(\kappa - 1)},$$

i.e. $\cos\alpha = \frac{2}{\kappa} - 1$. So $\alpha = \arccos(\frac{2}{\kappa} - 1)$ and $\sin\alpha = \sqrt{1-\cos^2\alpha} = 2\sqrt{\kappa - 1}/\kappa$.

(2) When $\kappa < 1$, C is a hyperbolic arc. Similar calculations to case (1) yield

$$S = Q_1 + \frac{1}{1-\cosh\alpha}[(Q_0 - Q_1) + (Q_2 - Q_1)].$$

i.e. $\cosh\alpha = \frac{2}{\kappa} - 1$. So $\alpha = \operatorname{arccosh}(\frac{2}{\kappa} - 1)$ and $\sinh\alpha = \sqrt{\cosh^2\alpha - 1} = 2\sqrt{1-\kappa}/\kappa$. □

Since only the magnitude of α, instead of its sign, matters in (2.30) and (2.31), α is always chosen to be positive in Theorem 2.4.2.

The application of Theorem 2.4.2 in drawing conic arcs is that, once the difference parameter k is determined from the precision requirement then (2.30) or (2.31) can be used to evaluate the initial points of the difference method on the conic arc. In order to determine the difference parameter k, according to Section 3, we have to analyze if arc C contains an apex of its underlying conic. The formulae for apices of the conic are given in [Lee87].

When the coordinates of the apices of the underlying conic have been obtained by the above formulae, we can use an inversion formula of the Bézier curve to determine the corresponding parameter values. An apex is on the conic arc if and only if the corresponding parameter value is in [0,1]. Suppose that the rational Bézier curve (2.28) is written in the form

$$x(t) = \frac{a_0 + a_1 t + a_2 t^2}{c_0 + c_1 t + c_2 t^2},$$
$$y(t) = \frac{b_0 + b_1 t + b_2 t^2}{c_0 + c_1 t + c_2 t^2}.$$

Then the following is one such inversion formula, given in [GSA84];

$$t = \frac{(b_0c_2 - b_2c_0)x + (a_2c_0 - a_0c_2)y + (a_0b_2 - a_2b_0)}{(b_2c_1 - b_1c_2)x + (a_1c_2 - a_2c_1)y + (a_2b_1 - a_1b_2)}.$$

There is also a direct inversion formula for conics in Bézier form in [GSA84].

Now we are able to give the following procedure for applying the difference method to a conic arc in the rational representation (2.28).

1. Compute the invariant $\kappa = w_0w_2/w_1^2$. If $\kappa \neq 1$, compute the center S from (2.29) and go to (2). If $\kappa = 1$, the arc C is a parabolic arc and (2.28) can be reparameterized to get all weights equal to 1 [Pat86], i.e. it can be put in the form (2.23). Then the difference method (2.12) with $k = 1$ can be applied directly. Determine the parameter increment θ by (2.25). The initial points are $P_0 = P(0)$, $P_1 = P(\theta)$ and $P_2 = P(2\theta)$.
2. If $\kappa > 1$, C is an elliptic arc; express it in the form (2.30), and compute parameter range $\alpha = \arccos(\frac{2}{\kappa} - 1)$. If $\kappa < 1$, C is a hyperbolic arc; express it in the form (2.31) and compute parameter range $\alpha = \operatorname{arccosh}(\frac{2}{\kappa} - 1)$.

3. For an elliptic arc (or a hyperbolic arc), compute the apices of the underlying conic and test whether any of them is on the arc. Determine $k = 2k'^2 - 1$ where k' satisfies (2.22) (or (2.20)). Compute the parameter increment $\theta = \arccos(k)$ (or $\operatorname{arccosh}(k)$).

4. Compute the initial points $P_0 = Q_0$, $P_1 = Q(\theta/\alpha)$ and $P_2 = Q(2\theta/\alpha)$ using (2.30) (or (2.31)). The number of points to be generated is given by $n = \lceil \alpha/\theta \rceil$. Then the difference formula (2.12) can be applied.

Note that the entire ellipse or the semi-ellipse cannot be expressed by (2.26) or (2.30), since $\sin\alpha \neq 0$. In these cases the arc can be first subdivided into smaller arcs in order to be represented by (2.30), then the above procedure can be applied. But when the expression for the elliptic arc introduced in the next subsection is used, such a subdivision is not necessary.

2.4.2 A new parameterization of elliptic arcs

In the last subsection we saw that the local parameterization (2.30) cannot represent the semi-ellipse, for the denominator of (2.30) $\sin\alpha = \sin\pi = 0$. In this subsection we shall find a new parameterization for the elliptic arc that includes the semi-ellipse, and at the same time uses the same parameter as in (2.30). This is achieved by using homogeneous coordinates. When the last component w of a homogeneous coordinate representation $P = [x, y, w]^T$ of a finite point equals one, the representation is said to be in the *normalized form.*

In this subsection it is assumed that an elliptic arc may be greater than the entire ellipse.

Theorem 2.4.3: *Let C be a circular arc on the unit circle $x^2 + y^2 = 1$. Let α be the signed central angle of the arc C; an arc with counterclockwise (CCW) direction has positive angle. Assume that $\alpha \neq 2l\pi$, l being an integer. Let $X_0 = [x_0, y_0, 1]^T$, $X_2 = [x_2, y_2, 1]^T$ be the homogeneous coordinates of the two endpoints of C. Then the arc has the parameterization*

$$X(v) = [1-\cos(1-v)\alpha]X_0 + \frac{[\sin(1-v)\alpha + \sin v\alpha - \sin\alpha]}{1-\cos\alpha} X_1 + [1-\cos(v\alpha)]X_2, \quad 0 \leq v \leq 1, \tag{2.32}$$

where

$$X_1 = [y_2 - y_0, x_0 - x_2, \sin\alpha]^T. \tag{2.33}$$

PROOF: Let the equation of the unit circle be $X^TAX = 0$, where X is a generic point in homogeneous coordinates, and $A = \mathrm{diag}[1, 1, -1]$. Since

$$\begin{aligned} X_0^TAX_1 &= x_0(y_2 - y_0) + y_0(x_0 - x_2) - \sin\alpha \\ &= x_0y_2 - x_2y_0 - \sin\alpha = \sin\alpha - \sin\alpha = 0, \end{aligned}$$

and similarly $X_2^TAX_1 = 0$, X_1 is the intersection of the two tangents $X_0^TAX = 0$ and $X_2^TAX = 0$ of the circle at X_0 and X_2. Therefore X_0, X_1 and X_2 can be regarded as the *control points* of the arc as in the Bézier representation.

By (2.26), the homogeneous representation of C is

$$X(v) = \begin{bmatrix} x_0 \sin(1-v)\alpha + x_2 \sin v\alpha \\ y_0 \sin(1-v)\alpha + y_2 \sin v\alpha \\ \sin\alpha \end{bmatrix} \tag{2.34}$$

Since $\alpha \neq 2l\pi$, l being an integer, X_0,X_1 and X_2 are linearly independent in homogeneous coordinates. So let

$$X(v) = a_0(v)X_0 + a_1(v)X_1 + a_2(v)X_2. \tag{2.35}$$

Then, since $X_0^TAX_0 = X_0^TAX_1 = X_2^TAX_2 = X_2^TAX_1 = 0$, we have

$$\begin{aligned} X(v)^TAX_0 &= a_2(v)X_2^TAX_0, \\ X(v)^TAX_1 &= a_1(v)X_1^TAX_1, \\ X(v)^TAX_2 &= a_0(v)X_0^TAX_2. \end{aligned}$$

The coefficients $a_0(v)$, $a_1(v)$ and $a_2(v)$ can be solved for from the above equations as follows. By (2.34)

$$\begin{aligned} X(v)^TAX_0 &= x_0[x_0 \sin(1-v)\alpha + x_2 \sin v\alpha] + y_0[y_0 \sin(1-v)\alpha + y_2 \sin v\alpha] - \sin\alpha \\ &= (x_0^2 + y_0^2)\sin(1-v)\alpha + (x_0x_2 + y_0y_2)\sin v\alpha - \sin\alpha \\ &= \sin(1-v)\alpha + \cos\alpha \sin v\alpha - \sin\alpha \\ &= \sin\alpha(\cos v\alpha - 1). \end{aligned}$$

Similarly, $X(v)^TAX_2 = \sin\alpha[\cos(1-v)\alpha - 1]$. And

$$\begin{aligned} X(v)^TAX_1 &= (y_2 - y_0)[x_0 \sin(1-v)\alpha + x_2 \sin v\alpha] \\ &\quad + (x_0 - x_2)[y_0 \sin(1-v)\alpha + y_2 \sin v\alpha] - \sin^2\alpha \\ &= (y_2x_0 - y_0x_2)\sin(1-v)\alpha + (x_0y_2 - x_2y_0)\sin v\alpha - \sin^2\alpha \\ &= \sin\alpha[\sin(1-v)\alpha + \sin v\alpha - \sin\alpha]. \end{aligned}$$

Also we have

$$X_0^T A X_2 = X_2^T A X_0 = x_0x_2 + y_0y_2 - 1 = \cos\alpha - 1,$$

and

$$\begin{aligned} X_1^T A X_1 &= (y_2 - y_0)^2 + (x_0 - x_2)^2 - \sin^2\alpha \\ &= x_0^2 + y_0^2 + x_2^2 + y_2^2 - 2(x_0x_2 + y_0y_2) - \sin^2\alpha \\ &= 2 - 2\cos\alpha - \sin^2\alpha = (1 - \cos\alpha)^2. \end{aligned}$$

So

$$\begin{aligned} a_0(v) &= \frac{X(v)^T A X_2}{X_0^T A X_2} = \frac{\sin\alpha[\cos(1-v)\alpha - 1]}{\cos\alpha - 1}, \\ a_1(v) &= \frac{X(v)^T A X_1}{X_1^T A X_1} = \frac{\sin\alpha[\sin(1-v)\alpha + \sin v\alpha - \sin\alpha]}{(1-\cos\alpha)^2}, \\ a_2(v) &= \frac{X(v)^T A X_0}{X_2^T A X_0} = \frac{\sin\alpha(\cos v\alpha - 1)}{\cos\alpha - 1}. \end{aligned}$$

Since we are using homogeneous coordinates, omitting a factor $\sin\alpha(1-\cos\alpha)^{-1}$, we can choose

$$\begin{aligned} a_0(v) &= 1 - \cos(1-v)\alpha, \\ a_1(v) &= \frac{[\sin(1-v)\alpha + \sin v\alpha - \sin\alpha]}{1 - \cos\alpha}, \\ a_2(v) &= 1 - \cos v\alpha. \end{aligned} \tag{2.36}$$

So (2.35) is the same as (2.32). Hence (2.32) represents a circular arc with $\alpha \neq 2l\pi$, l being an integer. □

When $\alpha = \pi$, $[x_2, y_2]^T = [-x_0, -y_0]^T$, it is easy to check that (2.32) reduces to

$$X(v) = \begin{bmatrix} x_0\cos v\pi - y_0\sin v\pi \\ y_0\cos v\pi + x_0\sin v\pi \\ 1 \end{bmatrix},$$

which represents a CCW semicircle starting at $[x_0, y_0, 1]^T$. When $\alpha = -\pi$ it can be similarly shown that (2.32) represents a clockwise semicircle starting at $[x_0, y_0, 1]^T$.

When the arc is a semicircle, the point X_1 is a point at infinity, so a point on the arc can be expressed as a linear combination of X_0, X_1 and X_2 in homogeneous coordinates. But in this case the two endpoints X_0 and X_2 are collinear with the origin, explaining why the representation (2.30) fails. Still a whole circle can not be represented by the new form (2.32), which in this case reduces to a point.

When the point X_1 is a finite point in the normalized form, (2.32) can be written as

$$X(v) = b_0(v)X_0 + b_1(v)X_1 + b_2(v)X_2, \tag{2.37}$$

where

$$\begin{aligned} b_0(v) &= 1 - \cos(1-v)\alpha, \\ b_1(v) &= \frac{\sin\alpha[\sin(1-v)\alpha + \sin v\alpha - \sin\alpha]}{1-\cos\alpha}, \\ b_2(v) &= 1 - \cos v\alpha. \end{aligned}$$

Now we generalize (2.32) to elliptic arcs.

Theorem 2.4.4: *An elliptic arc that is not the entire ellipse or a multiple of the ellipse can be represented in the form (2.32).*

PROOF: Consider an elliptic arc C which is not a multiple of an ellipse. Note that we are still using homogeneous coordinates. Let the endpoints of C be X_0 and X_2 in normalized form. Any elliptic arc can be mapped to a circular arc on the unit circle by an affine transformation. Let M be the nonsingular matrix of an affine mapping that maps C to a circular arc $\bar{C}$ on the unit circle such that the $\bar{X}_0 = MX_0$, $\bar{X}_2 = MX_2$ are also in normalized form. Then by Theorem 2.4.3, there exist α and $\bar{X}_1$ such that $\bar{C}$ is represented in the form (2.32), i.e.

$$\bar{X}(v) = a_0(v)\bar{X}_0 + a_1(v)\bar{X}_1 + a_2(v)\bar{X}_2,$$

where the $a_i(v)$ are given in (2.36). Applying M^{-1} to both sides of the above equation, and denoting $X_1 = M^{-1}\bar{X}_1$, we have the desired representation for the elliptic arc C. Since the last row of M and M^{-1} is $[0, 0, 1]$, the third component of X_1 is the same as that of $\bar{X}_1$. □

In order to apply Theorem 2.4.4 to an elliptic arc in the rational Bézier form (2.28), possibly with a nonpositive weight, one has to find the X_i and α. First X_0 and X_2 are the homogeneous coordinates of Q_0 and Q_2 in normalized form, and α can be determined from Theorem 2.4.2. In this case $0 < |\alpha| < 2\pi$ and the sign of α should be determined appropriately. When $\alpha \neq \pm\pi$, X_1 is the homogeneous coordinates of Q_1 with the last component being $\sin\alpha$. But when $\alpha = \pm\pi$, $\sin\alpha = 0$, i.e. the conic is a semi-ellipse and X_1 is a point at infinity. In this case we may find X_1 as follows.

To simplify the discussion, we assume that (2.28) is written in the standard form through the reparameterization [Pat86], i.e. the two end weights w_0 and w_2 are 1. As the conic is a semi-ellipse, we have the Bézier representation

$$P(t) = \frac{B_{0,2}(t)Q_0 + B_{1,2}(t)Q_1 + B_{2,2}(t)Q_2}{B_{0,2}(t) + B_{2,2}(t)}. \tag{2.38}$$

In this case we will show that $X_1 = [\pm 2Q_1^T, 0]^T$ when $\alpha = \pm\pi$, where Q_1 is a point in affine coordinates given in (2.38) and X_1 is in homogeneous coordinates. In the representation (2.32) or (2.38), a ***shoulder point*** of the curve is defined to be the point on the curve at which the tangent of the curve is parallel to the base $\overline{X_0X_2}$ or $\overline{Q_0Q_2}$. Since $\alpha = \pm\pi$, by simple algebra the shoulder point of the conic (2.32) can be shown to be

$$X(\frac{1}{2}) = X_0 \pm X_1 + X_2.$$

On the other hand, the shoulder point of the conic (2.38) can be shown to be

$$P(\frac{1}{2}) = \frac{1}{2}Q_0 + Q_1 + \frac{1}{2}Q_2.$$

Since these two points are the same, and considering that X_0 and X_2 are the homogeneous coordinates of Q_0 and Q_2 in normalized form, we have

$$X_1 = [\pm 2Q_1^T, 0]^T,$$

depending on $\alpha = \pm\pi$.

Next we consider the stability of using (2.32) to compute a point on the conic arc. By computing a point we mean to obtain the affine coordinates of the point. Therefore a stable evaluation requires that the third component of $X(v)$ be not very small, for otherwise division by a small number will greatly magnify the error in the numerator, resulting in large errors in the affine coordinates of the point.

The third component of $X(v)$ is

$$\begin{aligned}
x_3(v) &= [1-\cos(1-v)\alpha] + \frac{\sin\alpha}{1-\cos\alpha}[\sin(1-v)\alpha + \sin v\alpha - \sin\alpha] + [1-\cos v\alpha] \\
&= \frac{1}{1-\cos\alpha}\{(1-\cos\alpha)[2-\cos(1-v)\alpha - \cos v\alpha] \\
&\quad + \sin\alpha[\sin(1-v)\alpha + \sin v\alpha - \sin\alpha]\} \\
&= \frac{1}{1-\cos\alpha}[2\sin^2\frac{\alpha}{2}(2 - 2\cos\frac{\alpha}{2}\cos\frac{2v-1}{2}\alpha) \\
&\quad +2\sin\frac{\alpha}{2}\cos\frac{\alpha}{2}(2\sin\frac{\alpha}{2}\cos\frac{2v-1}{2}\alpha - \sin\alpha)] \\
&= \frac{1}{2\sin^2(\alpha/2)}\left(4\sin^2\frac{\alpha}{2} - 2\sin\frac{\alpha}{2}\cos\frac{\alpha}{2}\sin\alpha\right) \\
&= 2 - 2\cos^2\frac{\alpha}{2} = 1 - \cos\alpha.
\end{aligned}$$

Since $1-\cos\alpha = 0$ for $\alpha = 2n\pi$, $n = 0, \pm 1, \ldots$, the computation using (2.32) is not stable when the elliptic arc is very short or very close to a whole ellipse or a multiple of an ellipse.

A feature of (2.32) is that an elliptic arc that is greater than the whole ellipse can be represented, i.e. $|\alpha| > 2\pi$, but α is not an integer multiple of 2π. In this case, as v varies from 0 to 1, $X(v)$ traverses the arc.

Suppose the form (2.32) is given. If α is changed, then different elliptic curves will arise. So α plays a role similar to that played by the weights in the rational Bézier representation of conics. Complementary arcs can be represented easily when $-2\pi < \alpha < 2\pi$. One just has to replace α by $\alpha - 2\pi$ in (2.32) when $0 < \alpha < 2\pi$, and by $\alpha + 2\pi$ when $-2\pi < \alpha < 0$.

2.4.3 Conics in implicit form

In this subsection we discuss how to apply the difference method to draw a conic arc given by an implicit equation. A conic arc C in implicit form is specified by giving the implicit equation of the underlying conic, the two endpoints, and the drawing direction in the case of an elliptic arc. Let the equation of the conic be $f(x, y) = 0$, where $f(x, y)$ is a quadratic polynomial in x, y. Then this equation can be put into the form $X^T AX + 2b^T X + c = 0$, where $X = [x, y]^T$, and A is a 2×2 symmetric matrix. Arc C is elliptic, parabolic and hyperbolic according to $\det(A_1) > 0, = 0$, or < 0, respectively.

To apply the difference method we need to find the parameter range of the arc with respect to the canonical representation (2.2) or (2.4). This range determines the number of points to be generated on the arc once the difference parameter k is known. Also we need to find the apices of the underlying conic and test whether any of them is on the arc. This information is needed to determine an appropriate difference parameter k so that the generated polygon approximates C within the prescribed tolerance.

The next theorem gives a simple formula for computing the parameter range α for elliptic and hyperbolic arcs.

Theorem 2.4.5: *Let C be a conic arc on $X^T BX = 0$ which is an ellipse or hyperbola, where $X = [x, y, 1]$ and $B = \begin{bmatrix} A & b \\ b^T & c \end{bmatrix}$ is a 3×3 symmetric real matrix. Suppose the homogeneous coordinates of the two endpoints of C are $X_0 = [x_0, y_0, 1]^T$ and $X_2 = [x_2, y_2, 1]^T$. Let α be the parameter range of C with respect to the canonical parameterization (2.2) or (2.4). Then*

$$\cos\alpha = 1 - \frac{\det(A) X_0^T B X_2}{\det(B)}, \tag{2.39}$$

when C is elliptic; and

$$\cosh\alpha = 1 - \frac{\det(A)X_0^T B X_2}{\det(B)}, \tag{2.40}$$

when C is hyperbolic.

PROOF: We will only give the proof for the elliptic arc; the case for the hyperbolic arc follows almost identically.

Let $X^T BX = 0$ stand for the ellipse containing arc C. Since any ellipse is affinely equivalent to the unit circle, there exists a matrix $M = \begin{bmatrix} M_1 & h \\ 0 & 1 \end{bmatrix}$ representing an affine transformation such that $M^T BM = \mathrm{diag}[a, a, -a]$, $a \neq 0$. Obviously, M sends points on circle $x^2+y^2 = 1$ onto the arc C and keeps the normalized form of homogeneous coordinates of a point, i.e. the last component remains one. Let $\bar{X}_0 = M^{-1}X_0$, $\bar{X}_2 = M^{-1}X_2$, which are on $x^2 + y^2 = 1$. Then

$$\begin{aligned} \cos\alpha - 1 &= \bar{X}_0^T \mathrm{diag}[1, 1, -1]\bar{X}_2 = X_0^T (M^{-1})^T \mathrm{diag}[1, 1, -1]M^{-1}X_2 \\ &= \frac{X_0^T BX_2}{a}. \end{aligned}$$

Therefore

$$\cos\alpha = 1 + (X_0^T BX_2)/a. \tag{2.41}$$

Now let us compute the constant a. From $M^T BM = \mathrm{diag}[a, a, -a]$, $\det^2(M)\det(B) = -a^3$. On the other hand, since

$$\begin{aligned} M^T BM &= \begin{bmatrix} M_1^T & 0 \\ h^T & 1 \end{bmatrix} \begin{bmatrix} A & b \\ b^T & c \end{bmatrix} \begin{bmatrix} M_1 & h \\ 0 & 1 \end{bmatrix} \\ &= \begin{bmatrix} M_1^T AM_1 & M_1^T Ah + M_1^T b \\ h^T AM_1 + b^T M_1 & h^T Ah + 2h^T b + c \end{bmatrix}, \end{aligned}$$

we have $M_1^T AM_1 = \begin{bmatrix} a & 0 \\ 0 & a \end{bmatrix}$. It follows that $\det^2(M_1)\det(A) = a^2$. Since $\det(M) = \det(M_1)$,

$$a = \frac{a^3}{a^2} = -\frac{\det^2(M)\det(B)}{\det^2(M_1)\det(A)} = -\frac{\det(B)}{\det(A)}.$$

Substituting it in (2.41) yields

$$\cos\alpha = 1 - \frac{\det(A)X_0^T BX_2}{\det(B)}.$$

□

Here we omit the discussion about the computation of apices of a conic, since it can be found in standard textbooks on analytic geometry. Once the apices of the conic $f(x, y) = 0$ are determined it is straightforward to test whether or not an apex is on the arc C, upon which we will not elaborate. Then according to the discussion in Section 2.3 we can determine the difference parameter k to satisfy a prescribed approximation tolerance. The next step is to determine the initial points for applying the difference method.

Suppose the equation of the underlying conic is $f(x, y) \equiv Y^T AY + 2b^T Y + c = 0$, where $Y = [x, y]^T$ are affine coordinates. First assume that C is not a parabolic arc. In this case, since $\det(A) \neq 0$, it is easy to verify that

$$f(x, y) = (Y - S)^T A(Y - S) + d, \tag{2.42}$$

where $S = -A^{-1}b$, and $d = c - b^T A^{-1}b$. So, assuming $d \neq 0$ for otherwise the conic is degenerate, the equation $f(x, y) = 0$ is equivalent to

$$(Y - S)^T(-A/d)(Y - S) = 1.$$

Let the second initial point be P_1. The third point, by (2.6), is

$$P_2 = 2k(P_1 - S) - (P_0 - S) + S. \tag{2.43}$$

Then we have

$$(P_1 - S)^T A(P_1 - S) + d = 0,$$

and

$$(P_2 - S)^T A(P_2 - S) + d = 0,$$

or, after substituting (2.43) in the last equation, and considering that P_0 and P_1 satisfy (2.42), we have

$$-4k(P_0 - S)^T A(P_1 - S) - 4k^2 d = 0,$$

or, assuming $k \neq 0$,

$$(P_0 - S)^T A(P_1 - S) + kd = 0.$$

When $k = 0$, C is on an elliptic arc and $\alpha = \pm\pi/2$. In this case $\overline{P_0S}$ and $\overline{P_1S}$ are on conjugate diameters of $f(x, y) = 0$, that is,

$$(P_0 - S)^T A(P_1 - S) = 0.$$

Therefore, in all cases, the initial point P_1 is determined by the system

$$\begin{cases} (P_1 - S)^T A(P_1 - S) + d = 0 \\ (P_0 - S)^T A(P_1 - S) + kd = 0. \end{cases}$$

Geometrically, the solutions of this system are the intersections of the conic $f(x, y) = 0$ with a straight line. The two solutions stand for the two directions to draw the conic arc C; the one that is consistent with the specified drawing direction should be chosen when arc C is elliptic.

When C is given by an implicit equation known to be parabolic, it can be readily expressed as a quadratic Bézier curve by finding the middle control point as the intersection of the two tangents of C at its endpoints. Then the approach in Section 2.4.1 can be used.

2.5 More properties of the difference method

2.5.1 A best approximation property

The best approximation property we shall discuss below is stated in [Smi71]. But the proof given there is incorrect as we shall explain below. So we now give a proof of this property.

Theorem 2.5.1: *Let C be a conic arc with fixed endpoints P_0 and P_n and $\{P_i\}_{i=0}^{n}$ be a point sequence on C with the P_i's, $i = 1, 2, \ldots, n-1$, being variable points on C having increasing arclength order. Let $\mathcal{G}_n$ be the set of all $(n+1)$-sided convex polygons $P_0P_1..P_n$ obtained by connecting P_i and P_{i+1}, $i = 0, 1, \ldots, n-1$, and connecting P_n and P_0. Let $S(G)$ denote the area of convex polygon G. Then* $\max\{S(G) | G \in \mathcal{G}_n\}$ *exists and it is attained if and only if $\{P_i\}_{i=0}^{n}$ satisfies the difference equation (2.12).*

To prove this theorem we need the following lemma.

Lemma 2.5.2: *Let C be a conic arc with fixed endpoints P_0 and P_n and $\{P_i\}_{i=0}^{n}$, $n \geq 2$, be a point sequence on C with increasing arclength order. Let T_i be the tangent of C at P_i. Then $\{P_i\}_{i=0}^{n}$ is a point sequence generated by the difference method if and only if T_i is parallel to the segment $\overline{P_{i-1}P_{i+1}}$, $i = 1, 2, \ldots, n-1$.*

PROOF: We will only consider the case of hyperbolas. Proofs for ellipses and parabolas are similar.

Necessity: Let C be a hyperbolic arc. Since the property to be proved is invariant under affine transformations, we assume that the equation of C is given by (2.4). By Theorem 2.2.1,

$$P_i = [\cosh(t_0 + i\theta), \sinh(t_0 + i\theta)]^T, \quad i \geq 0,$$

for some $\theta \neq 0$. The tangent of C at $P(t)$ is

$$P'(t) = [\sinh t,\ \cosh t]^T.$$

So

$$T_i = P'(t)|_{t=t_0+i\theta} = [\sinh(t_0+i\theta),\ \cosh(t_0+i\theta)]^T.$$

On the other hand, the vector

$$\begin{aligned}\overrightarrow{P_{i-1}P_{i+1}} &= [\cosh(t_0+(i+1)\theta), \sinh(t_0+(i+1)\theta)]^T \\ &\quad -[\cosh(t_0+(i-1)\theta), \sinh(t_0+(i-1)\theta)]^T \\ &= [2\sinh\theta\sinh(t_0+i\theta),\ 2\sinh\theta\cosh(t_0+i\theta)]^T \\ &= 2\sinh\theta\, T_i.\end{aligned}$$

Hence T_i is parallel to line segment $\overline{P_{i-1}P_{i+1}}$.

Sufficiency: Assume again that the hyperbolic arc C has the canonical representation $P(t)$ in (2.4). Let $P_i = P(t_i)$ and $t_0 \leq t_1 \leq \ldots \leq t_{n-1} \leq t_n$, where t_0 and t_n are fixed. Then it is easy to show that

$$\overrightarrow{P_{i-1}P_{i+1}} = 2\sinh\left(\frac{t_{i+1}-t_{i-1}}{2}\right)\left[\sinh\left(\frac{t_{i+1}+t_{i-1}}{2}\right),\ \cosh\left(\frac{t_{i+1}+t_{i-1}}{2}\right)\right]^T,$$

and

$$T_i = [\sinh t_i,\ \cosh t_i]^T,$$

Since T_i is parallel to $\overrightarrow{P_{i-1}P_{i+1}}$, we have

$$\frac{\sinh t_i}{\cosh t_i} = \frac{\sinh[(t_{i+1}+t_{i-1})/2]}{\cosh[(t_{i+1}+t_{i-1})/2]}.$$

So

$$\begin{aligned}&\sinh[(t_{i+1}+t_{i-1})/2]\cosh t_i - \cosh[(t_{i+1}+t_{i-1})/2]\sinh t_i \\ &= \sinh\left(\frac{t_{i+1}+t_{i-1}}{2} - t_i\right) = 0, \quad i = 1, 2, \ldots, n-1,\end{aligned}$$

i.e. $t_i = (t_{i+1}+t_{i-1})/2$. Thus there exists $\theta = (t_n - t_0)/n$ such that $t_i = t_0 + i\theta$. Hence $\{P_i\}_{i=0}^n$ satisfies the difference equation (2.12). □

Proof of Theorem 2.5.1: First we show that the maximum of $S(G)$ exists in the set $\mathcal{G}_n$ of all convex polygons defined above. Let the equation of the conic arc C be $P(t)$ which is in the appropriate canonical form (2.2), (2.4) or (2.23). Let $P_i = P(t_i)$. Then, by assumption, $t_0 \leq t_1 \leq \ldots \leq t_{n-1} \leq t_n$. Define $M = \{(t_1, t_2, \ldots, t_{n-1}) | t_0 \leq t_1 \leq \ldots t_{n-1} \leq t_n\}$ in E^{n-1}. Obviously M is a compact set in E^{n-1}.

Let G be a convex polygon with vertices P_i, $i = 0, 1, \ldots, n$. Then the area $S(G)$ is a continuous function of P_i, therefore it is a continuous function of t_i, $i = 1, 2, \ldots, n-1$, since $P(t)$ is continuous. So $S(G)$ attains a maximum in the compact set M, and hence it has a maximum in $\mathcal{G}_n$.

Consider the necessity part first. Suppose for polygon G with vertices P_0, $P_1, \ldots,$ P_n, $S(G)$ attains the maximum in $\mathcal{G}_n$. Then we claim that the tangent T_i of arc C at P_i is parallel to the segment $\overline{P_{i-1}P_{i+1}}$, $i = 1, 2, \ldots, n-1$. First of all, we see that no two consecutive points P_i and P_{i+1} can be identical, for otherwise we can always find three consecutive points P_j, P_{j+1} and P_{j+2} such that $P_j = P_{j+1} \neq P_{j+2}$ or $P_j \neq P_{j+1} = P_{j+2}$ since not all points are identical; without loss of generality, we assume the former case holds. Then moving P_{j+1} away from P_j towards P_{j+2} will increase the area of polygon G since the arc C is convex. But this contradicts that $S(G)$ is maximum.

So we have shown that all the P_i's are distinct. Suppose that the tangent T_i of C at P_i is not parallel to $\overline{P_{i-1}P_{i+1}}$ for some i. Then one can move the the point P_i to the point $\bar{P}_i$ on arc $P_{i-1}P_{i+1}$ that yields the maximum distance to the segment $\overline{P_{i-1}P_{i+1}}$. But this would increase the area of $S(G)$, contradicting that $S(G)$ is maximum. So T_i is parallel to $\overline{P_{i-1}P_{i+1}}$ for all $i = 1, 2, \ldots, n-1$. Hence, by Lemma 2.5.2, the point sequence $\{P_i\}_{i=0}^{n}$ satisfies the difference equation (2.12).

Now consider the sufficiency part. Suppose $\{P_i\}_{i=0}^{n}$ satisfy the difference equation (2.12). Now we need to be more specific about the arc C. We will assume that C is a hyperbolic arc and it is in the form (2.4) . The other two cases can be treated similarly. Let S_i denote the area of the region enclosed by the arc P_iP_{i+1} and the segment $\overline{P_iP_{i+1}}$. Since $\sum_{i=0}^{n-1} S_i$ is the area of the region between the arc C and the polygonal line $P_0P_1 \cdots P_n$, $S(G)$ attains the maximum if and only if $F(G) \equiv \sum_{i=0}^{n-1} S_i$ is minimum. Therefore, we just have to show that $F(G)$ is minimum when G's vertices satisfy (2.12).

By the calculation as given in the proof of Lemma 2.3.2, we have $S_i = \frac{1}{2}[\sinh(t_{i+1} - t_i) - (t_{i+1} - t_i)]$. So

$$F = \frac{1}{2}\sum_{i=0}^{n-1} \sinh(t_{i+1} - t_i) - (t_n - t_0).$$

Therefore

$$\begin{aligned} \frac{\partial F}{\partial t_i} &= \frac{1}{2}[\cosh(t_i - t_{i-1}) - \cosh(t_{i+1} - t_i)] \\ &= -\sinh\frac{t_{i+1} - t_{i-1}}{2}\sinh\frac{2t_i - t_{i-1} - t_{i+1}}{2}, \quad i = 1, 2, \ldots, n-1. \end{aligned}$$

One consequence of the preceding necessity proof is that all the P_i's are distinct when

$S(G)$ is maximum. Recalling the compact set M defined above, the boundary of M is characterized by $t_j = t_{j+1}$ for some $0 \leq j < n$, i.e. there are identical points in the P_i's. So the maximum of $S(G)$ can only be attained in $In(M)$, the interior of M. Therefore the system $\partial F/\partial t_i = 0$, $i = 1, 2, \ldots, n-1$, has a solution in $In(M)$ which gives the maximum $S(G)$, or minimum $F(G)$, $G \in \mathcal{G}_n$.

On the other hand, it is easy to see that the system $\partial F/\partial t_i = 0$, $i = 1, 2, \ldots, n-1$, has a unique solution in $In(M)$, which is,

$$t_i = \frac{(t_{i-1} + t_{i+1})}{2}, \quad i = 1, 2, \ldots, n-1.$$

So this is the condition for polygon $P(t_0)P(t_1)\cdots P(t_n)$ to have the maximum area. But by Theorem 2.2.1, this condition is satisfied by the t_i's corresponding to the sequence $\{P_i\}_{i=0}^{n}$ which satisfies the difference equation (2.12). So $S(G)$ is maximum over $\mathcal{G}_n$ when its vertices are generated by the difference method. □

Now we explain why the proof given in [Smi71] for Theorem 2.5.1 is incorrect. It is argued there that for a point sequence $\{P_i\}_{i=0}^{n}$ with increasing arclength order on a smooth convex arc, with the two endpoints fixed, if the tangent T_i of C at P_i is parallel to the segment $\overline{P_{i-1}P_{i+1}}$, $i = 1, 2, \ldots, n-1$, then the area of the convex polygon $P_0P_1\cdots P_n$ is maximum. In fact, from the given condition one can only deduce that the polygon has locally maximal area, but, in general, not globally maximal area. One counterexample is shown in Fig. 2.5.1. The arc shown is part of a curve that is obtained by bulging a square slightly, making it smooth and symmetric. In Fig. 2.5.1(a), the points P_i, $i = 0, 1, 2, 3$, satisfy the condition that the tangent at P_i is parallel to $\overline{P_{i-1}P_{i+1}}$, $i = 1, 2$; however its area is smaller than that of the polygon $P_0P_1P_2P_3$ depicted in Fig. 2.5.1(b).

2.5.2 A projective property

In this subsection we study a projective property of a point sequence generated by the difference method. From this property we can answer the following question: Can a point sequence generated on a hyperbola or an ellipse also be generated by the forward differencing method? Or, more generally, when can two point sequences generated by the difference method be mapped onto each other by a projective transformation?

First we need some notation about the projective geometry of a conic section. The following lemma enables us to speak about the cross ratio of four points on a conic.

Lemma 2.5.3: *Let P_0, P_1, P_2 and P_3 be four points on a conic C. Then for any point*

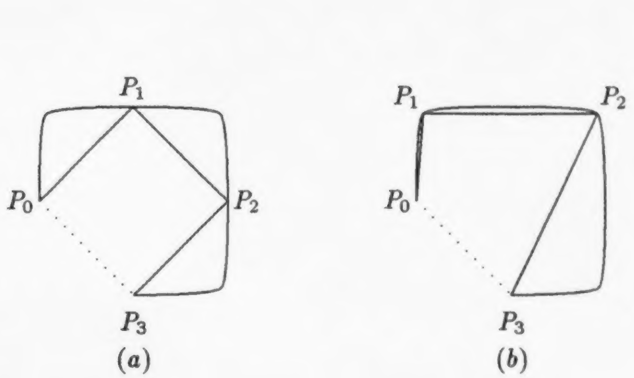


Figure 2.5.1 A counterexample: Two polygonal approximations of a smooth convex curve which is part of a slightly bulging square. If $|P_0P_2| = 1$ in (a), then the area of closed polygon $P_0P_1P_2P_3$ in (a) is about 1/2, while that of $P_0P_1P_2P_3$ in (b) is about 5/8.

K on C distinct from any P_i, $i = 0, 1, 2, 3$, the four lines KP_0, KP_1, KP_2 and KP_3 have the same cross ratio.

This lemma is known as Chasles's Theorem, and its proof can be found in [SeK50, p. 133]. The cross ratio of four lines l_i, $i = 0, 1, 2, 3$, concurrent at point K in a plane, is denoted by $(l_0, l_2; l_1, l_3)$, which is defined to be the cross ratio $(Q_0, Q_2; Q_1, Q_3)$ of the four points Q_i on the l_i intersected by a line l not passing through K. For four collinear points Q_i, $i = 0, 1, 2, 3$, the cross ratio $(Q_0, Q_2; Q_1, Q_3)$ is defined by

$$\frac{\overrightarrow{Q_0Q_1}}{\overrightarrow{Q_2Q_1}} \Big/ \frac{\overrightarrow{Q_0Q_3}}{\overrightarrow{Q_2Q_3}},$$

where $\overrightarrow{Q_iQ_j}$ is the directed length of segment $\overline{Q_iQ_j}$. This definition for $(l_0, l_2; l_1, l_3)$ makes sense since it is a basic result of projective geometry that the quantity thus obtained does not depend on any particular choice of l. Now we can define the cross ratio of four points P_i, $i = 0, 1, 2, 3$, on a conic.

Definition 2.5.1: *Let P_i, $i = 0, 1, 2, 3$, be four points on a proper conic C. The cross ratio of the P_i, $i = 0, 1, 2, 3$, with respect to C is*

$$(P_0, P_2; P_1, P_3)_C = (KP_0, KP_2; KP_1, KP_3)_C,$$

where K is a point on C which is not equal to any of the P_i but otherwise arbitrary, and KP_i denotes the line passing through K and P_i, $i = 0, 1, 2, 3$.

The subscript C indicates that the cross ratio is defined with respect to the conic C, and it will be omitted if the context makes the reference to the conic clear. It is important to note that by Lemma 2.5.3 the cross ratio thus defined does not depend on any particular choice of K on C. Hence on every conic we have defined a one dimensional projective geometry induced by the projective geometry of the plane.

Since the cross ratio of four concurrent lines in the plane is invariant under projective transformations, we have the following obvious lemma.

Lemma 2.5.4: *Let P_0, P_1, P_2 and P_3 be four points on a conic C. Let M be a projective transformation that maps conic C to conic C', and the P_i's to the P_i''s on C', $i = 0, 1, 2, 3$, respectively. Then*

$$(P_0, P_2;\ P_1, P_3)_C = (P_0', P_2';\ P_1', P_3')_{C'}.$$

Definition 2.5.2: *Let $\{P_i\}_{i \geq 0}$ be a point sequence on conic C. If $(P_i, P_{i+2};\ P_{i+1}, P_{i+3})$ is the same for all $i \geq 0$, then $\{P_i\}_{i \geq 0}$ is called a* projective sequence *with cross ratio $r = (P_0, P_2;\ P_1, P_3)_C$.*

We will assume that any projective sequence lies on a unique conic; this can be guaranteed by insisting that the sequence has at least five distinct points in general position. Therefore a projective sequence has a unique cross ratio.

Theorem 2.5.5: *Let $\{P_i\}_{i \geq 0}$ be a point sequence satisfying the difference equation (2.12) with difference parameter k. Then $\{P_i\}_{i \geq 0}$ is a projective sequence with cross ratio $r = -1/(2k+1)$.*

PROOF: We will consider just the case of hyperbolas; the cases of ellipses and parabolas can be proved similarly. Since the cross ratio is certainly not altered by affine transformations, assume that C has the representation $P(t) = [\cosh t,\ \sinh t]^T$. As $\{P_i\}_{i \geq 0}$ satisfies the difference equation (2.12), we can assume $P_i = P(t_0 + i\theta)$, $i \geq 0$, for some t_0 and $\theta \neq 0$.

We first compute the cross ratio of P_1, P_2, P_3 and P_4 on C. Consider four straight lines P_0P_i, $i = 1, 2, 3, 4$. The direction vector of line P_0P_i is

$$P_i - P_0 = 2\sinh(i\theta/2)[\sinh(t_0 + i\theta/2),\ \cosh(t_0 + i\theta/2)]^T.$$

So the equation of line P_0P_i is

$$l_i \equiv \cosh(t_0 + i\theta/2)(x - x_0) - \sinh(t_0 + i\theta/2)(y - y_0) = 0, \quad i = 1, 2, 3, 4,$$

where $[x_0, y_0]^T = P_0$. Since $\sinh(t_0 + i\theta/2)$ and $\cosh(t_0 + i\theta/2)$ satisfy the difference equation

$$x_{i+2} = 2k'x_{i+1} - x_i, \quad i \geq 0,$$

where $k' = \cosh(\theta/2)$, we have

$$l_3 = 2k'l_2 - l_1$$

and

$$l_4 = 2k'l_3 - l_2.$$

Substituting $l_3 = 2k'l_2 - l_1$ in the above expression,

$$l_4 = (4k'^2 - 1)l_2 - 2k'l_1.$$

Consider now the one dimensional projective space of the pencil of lines passing through P_0. We can establish a projective coordinate system of this space so that the projective coordinates $(\lambda_i,\ \mu_i)$ of l_i, $i = 1,2,3,4$, are $(0,\ 1)$, $(1,\ 0)$, $(2k', -1)$ and $(4k'^2 - 1,\ -2k')$, respectively. Then, by the definition of the cross ratio [SeK50, p. 25],

$$\begin{aligned}(P_0P_1, P_0P_3;\ P_0P_2, P_0P_4) &= \frac{(\lambda_1\mu_2 - \lambda_2\mu_1)(\lambda_3\mu_4 - \lambda_4\mu_3)}{(\lambda_3\mu_2 - \lambda_2\mu_3)(\lambda_1\mu_4 - \lambda_4\mu_1)} \\ &= \frac{-1}{4k'^2 - 1} = \frac{-1}{4\cosh^2(\theta/2) - 1} \\ &= \frac{-1}{2\cosh\theta + 1} = \frac{-1}{2k+1}.\end{aligned}$$

So, by Definition 2.5.1,

$$(P_1, P_3;\ P_2, P_4) = (P_0P_1, P_0P_3;\ P_0P_2, P_0P_4) = \frac{-1}{2k+1}.$$

As the above derivation can be repeated for any four consecutive points P_i, P_{i+1}, P_{i+2} and P_{i+3} with P_0 replaced by P_{i-1}, we have

$$(P_i, P_{i+2};\ P_{i+1}, P_{i+3}) = \frac{-1}{2k+1}, \quad i \geq 0.$$

□

Corollary 2.5.6: *The point sequence $\{P_i\}_{i\geq 0}$ generated by the difference method on an ellipse or a hyperbola C cannot be reproduced by the forward differencing method.*

PROOF: According to Theorem 2.2.2, $\{P_i\}_{i\geq 0}$ is generated with difference parameter $k \neq 1$. By Theorem 2.5.5, $\{P_i\}_{i=0}^{n}$ is a projective sequence with cross ratio $-1/(2k+1) \neq -1/3$. But any point sequence $\{Q_i\}_{i\geq 0}$ on C generated by forward differencing is the image under a projective transformation of a point sequence $\{Q'_i\}_{i\geq 0}$ on a parabola that is generated by the difference method with $k = 1$. By Theorem 2.5.5, $\{Q'_i\}_{i\geq 0}$ is a projective sequence with cross ratio $-1/3$. So, by Lemma 2.5.4, the cross ratio of $\{Q_i\}_{i\geq 0}$ is also $-1/3$. Hence $\{Q_i\}_{i\geq 0}$ and $\{P_i\}_{i\geq 0}$ cannot be the same sequence since they have different cross ratios. □

2.5.3 On generalization

In this subsection we consider whether the difference equation (2.12) can be generalized to generate point sequences on rational cubic curves. This question suggests itself naturally in the following observation. It is well known that the difference equation

$$P_{i+3} = 3P_{i+2} - 3P_{i+1} + P_i, \quad i \geq 0,$$

generates points on a parabola, which is a polynomial quadratic curve. By generalizing it to the equation

$$P_{i+3} = (2k+1)P_{i+2} - (2k+1)P_{i+1} + P_i, \quad i \geq 0,$$

we have shown that all conics, i.e. all rational quadratic curves can be generated. Now, since

$$P_{i+4} = 4P_{i+3} - 6P_{i+2} + 4P_{i+1} - P_i, \quad i \geq 0,$$

can generate all polynomial cubic curves, we ask whether for any rational cubic curve $P(t)$ one can find a generalization of this equation, which is in the form,

$$P_{i+4} = c_3 P_{i+3} + c_2 P_{i+2} + c_1 P_{i+1} + c_0 P_i, \quad i \geq 0, \tag{2.44}$$

where the c_i are real, so that a point sequence on $P(t)$ can be generated.

We will show that such a generalization is impossible by proving that there exist some rational cubic curves such that there is no fourth order difference equation that can generate point sequences on them. If $\{P_i\}_{i\geq 0}$ can be generated on a curve $P(t)$, we assume that, by the natural extension of $\{P_i\}_{i\geq 0}$ determined by (2.44), the point sequence $\{P_n\}_{-\infty}^{\infty}$ is also on $P(t)$, i.e. there is a point sequence $\{P_n\}_{-\infty}^{\infty}$ on $P(t)$ that satisfies Eqn. (2.44). Since a rational cubic curve is always unbounded in E^3, we assume that a point sequence generated on a cubic curve by the difference method is also unbounded if such generation is possible.

Theorem 2.5.7: *The fourth order difference equation cannot generate an unbounded point sequence on some rational cubic curves.*

PROOF: We just need to show that an equation of form (2.44) cannot generate the rational cubic space curve

$$P(t) = [x(t), y(t), z(t)]^T = \left[\frac{1-t^2}{t^2+1}, \frac{2t}{t^2+1}, t\right]^T, \quad -\infty < t < \infty. \tag{2.45}$$

The proof is by contradiction. So suppose that $\{P_n\}_{-\infty}^{\infty}$ is an unbounded point sequence on the curve $P(t)$ and it satisfies (2.44) for some constant c_j's. Let

$$P_n = [x_n, y_n, z_n]^T = P(t_n) = [x(t_n), y(t_n), z(t_n)]^T, \quad -\infty < n < \infty. \tag{2.46}$$

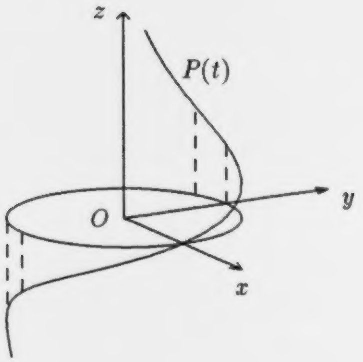


Figure 2.5.2 The cubic curve $P(t)$ and its projection.

Since the parallel projection of $P(t)$ onto the xy-plane along the z-axis is the unit circle $x^2 + y^2 = 1$, the sequence $\{[x_n, y_n]^T\}_{-\infty}^{\infty}$ lies on this circle. (See Fig. 2.5.2.) Moreover, the x_n's and y_n's still satisfy (2.44), which is understood as an equation in scalars instead of in vectors. Now the sequence $\{z_n\}_{-\infty}^{\infty}$ has to be unbounded since $\{P_n\}_{-\infty}^{\infty}$ is unbounded.

For a difference equation

$$\sum_{j=0}^{m} a_j P_{i+j} = 0,$$

its characteristic equation is given by

$$\sum_{j=0}^{m} a_j s^j = 0.$$

The solutions P_n of the difference equation are linear combinations of *n-th* powers of the roots of its characteristic equation when all the roots are distinct; when a root s_0 has multiplicity p, the coefficient of s_0^n in P_n is a polynomial of order p in n.

So the characteristic equation of (2.44) is

$$s^4 - c_3 s^3 - c_2 s^2 - c_1 s - c_0 = 0. \tag{2.47}$$

Let the four roots of this equation be s_j, $j = 1, 2, 3, 4$. First assume that these roots are distinct. Then

$$[x_n, y_n]^T = A_1 s_1^n + A_2 s_2^n + A_3 s_3^n + A_4 s_4^n, \tag{2.48}$$

and

$$z_n = e_1 s_1^n + e_2 s_2^n + e_3 s_3^n + e_4 s_4^n, \tag{2.49}$$

where the A_j and e_j are constants, $j = 1, 2, 3, 4$. These two expressions are to be appropriately modified when there are multiple roots in the s_j, e.g. if $s_2 = s_1$ is a double root then s_2^n is replaced with ns_1^n.

For any $j = 1, 2, 3$, or 4, if $A_j \neq 0$, then we can show that $|s_j| = 1$, i.e. they are unit roots. For otherwise, assuming $A_1 \neq 0$, $|s_1| < 1$ and $|s_1|$ is the smallest among all the $|s_j|$'s whose coefficient $A_j \neq 0$, then the right hand side of (2.48) will tend to infinity as $n \to -\infty$. This contradicts the fact that the $[x_n, y_n]^T$'s are on the unit circle. Similarly it can be shown that $|s_j|$ can not be greater than 1 by considering $n \to +\infty$. Therefore in (2.48) $[x_n, y_n]^T$ is represented as a linear combination of powers of unit roots of Eqn. (2.47).

In the case of repeated roots, it can be similarly shown that all roots that appear in (2.48) with a nonzero coefficient must also be unit roots. Actually, no terms that contains a subsequence tending to infinity are allowable in the right hand of (2.48).

Since there are an infinite number of distinct points of $\{P_n\}_{-\infty}^{\infty}$ on $P(t)$ and since different points project to different points on the unit circle, there is at least one pair of conjugate complex unit roots, say, $s_{1,2} = e^{\pm i\theta}$ in the s_j's, for the only real unit roots 1 and -1 can generate at most two distinct points on the circle $x^2 + y^2 = 1$. So

$$\begin{aligned} [x_n, y_n]^T &= A_1 s_1^n + A_2 \bar{s}_1^n + A_3 s_3^n + A_4 s_4^n \\ &= C_1 \cos(n\theta) + C_2 \sin(n\theta) + A_3 s_3^n + A_4 s_4^n. \end{aligned}$$

Now we shall show that $A_3 = A_4 = 0$. Suppose that $A_3 \neq 0$ or $A_4 \neq 0$. From the preceding discussion we must have $|s_3| = |s_4| = 1$, and s_3 and s_4 can not be a conjugate pair that is the same as s_1 and s_2. But when s_3 and s_4 is a complex conjugate pair that is different from s_1 and s_2, a contradiction arises from (2.49) since the left hand side of (2.49) is unbounded but the right hand side is bounded. A similar argument shows that s_3 and s_4 can not be distinct real roots, i.e. $s_3 = -1$ and $s_4 = 1$. Now suppose that $s_3 = s_4 = 1$. Due to the boundedness of $[x_n, y_n]^T$, the coefficients of s_3 and s_4 can not be both nonzero. So we can assume that $A_4 = 0$ and $A_3 \neq 0$. Therefore

$$[x_n, y_n]^T = C_1 \cos(n\theta) + C_2 \sin(n\theta) + A_3,$$

which is a point sequence on an ellipse with its center A_3 not at the origin. Since we have infinite distinct points on this ellipse and since this ellipse can have at most four points of intersection with unit circle $x^2 + y^2 = 1$ on the $z = 0$ plane, we conclude that the point sequence can not be on the unit circle. This is a contradiction. Similarly, a contradiction follows from the case that $s_3 = s_4 = -1$. Hence we must have $A_3 = A_4 = 0$, i.e.

$$[x_n, y_n]^T = C_1 \cos(n\theta) + C_2 \sin(n\theta), \quad -\infty < n < \infty. \tag{2.50}$$

Now we claim that C_1 and C_2 must be real and orthonormal vectors. First, from the above equation we have

$$[x_n, y_n]^T = \bar{C}_1 \cos(n\theta) + \bar{C}_2 \sin(n\theta), \quad -\infty < n < \infty,$$

where $\bar{C}_1$ and $\bar{C}_2$ are the complex conjugates of C_1 and C_2. Then, subtracting this equation from the last one, we have

$$(C_1 - \bar{C}_1)\cos(n\theta) + (C_2 - \bar{C}_2)\sin(n\theta) = 0, \quad -\infty < n < \infty.$$

Since θ is such that infinitely many distinct points are generated by (2.50), the sequence $\{\cos(n\theta)\}_{-\infty}^{\infty}$ and $\{\sin(n\theta)\}_{-\infty}^{\infty}$ are linearly independent, i.e. one can not be expressed as a linear multiple of the other. Then it follows that $C_1 = \bar{C}_1$ and $C_2 = \bar{C}_2$, that is, C_1 and C_2 are real. By considering the condition $[x_n, y_n][x_n, y_n]^T = 1$, it is easy to show that $C_1^T C_1 = C_2^T C_2 = 1$ and $C_1^T C_2 = 0$. Therefore C_1 and C_2 are real orthonormal vectors.

Let $C_1 = [\cos\alpha, \sin\alpha]^T$ and $C_2 = [-\sin\alpha, \cos\alpha]^T$ for some constant α. Then

$$[x_n, y_n]^T = [\cos(\alpha + n\theta),\ \sin(\alpha + n\theta)]^T.$$

But as

$$[x_n, y_n]^T = \left[\frac{1 - t_n^2}{t_n^2 + 1}, \frac{2t_n}{t_n^2 + 1}\right]^T,$$

we obtain

$$t_n = \tan(\frac{\alpha + n\theta}{2}).$$

Thus

$$z_n = \tan(\frac{\alpha + n\theta}{2}), \quad -\infty < n < \infty, \tag{2.51}$$

since $z(t) = t$ by (2.45).

On the other hand, different values of s_3 and s_4 would generate different expressions for z_n through (2.49) and its modified form in the case of multiple roots, and consequently different expressions for $\tan[(\alpha + n\theta)/2]$ by (2.51). Since any such expression for z_n must be unbounded, in order for the right hand side also to be unbounded, there are the following four possible cases to consider.

(1) $s_3 = s_1$ and $s_4 = s_2$. In this case there are a pair of repeated conjugate roots and

$$\tan(\frac{\alpha + n\theta}{2}) = d_1 \cos(n\theta) + d_2 \sin(n\theta) + n[d_3 \cos(n\theta) + d_4 \sin(n\theta))],$$

where $d_3 \neq 0$ or $d_4 \neq 0$.

(2) s_3 and s_4 are conjugate, but distinct from the pair s_1 and s_2. Let $s_3 = \bar{s}_4 = \rho e^{i\beta}$, and $\rho \neq 1$. In this case

$$\tan(\frac{\alpha + n\theta}{2}) = d_1 \cos(n\theta) + d_2 \sin(n\theta) + \rho^n [d_3 \cos(n\beta) + d_4 \sin(n\beta)],$$

where $d_3 \neq 0$ or $d_4 \neq 0$.

(3) $s_3 = s_4 = \rho$ are real. In this case

$$\tan(\frac{\alpha + n\theta}{2}) = d_1 \cos(n\theta) + d_2 \sin(n\theta) + d_3 \rho^n + n d_4 \rho^n,$$

where $d_4 \neq 0$ if $|\rho| = 1$, otherwise $d_3 \neq 0$ or $d_4 \neq 0$.

(4) s_3 and s_4 are real, $s_3 \neq s_4$, and $|s_3| \neq 1$ or $|s_4| \neq 1$.

$$\tan(\frac{\alpha + n\theta}{2}) = d_1 \cos(n\theta) + d_2 \sin(n\theta) + d_3 s_3^n + d_4 s_4^n,$$

where $d_3 \neq 0$ if $|s_4| = 1$, $d_4 \neq 0$ if $|s_3| = 1$, otherwise $d_3 \neq 0$ or $d_4 \neq 0$.

Now we show that none of the above cases can hold. This contradiction will complete the proof. In case (1), we have

$$\frac{\tan[(\alpha + n\theta)/2] - d_1 \cos(n\theta) - d_2 \sin(n\theta)}{d_3 \cos(n\theta) + d_4 \sin(n\theta)} = n.$$

The left hand side is a periodic function of θ, so we can always choose a subsequence $0 < n_1 < n_2 < \ldots < +\infty$ such that the left hand side remains bounded. But for any such subsequence the right side tends to $+\infty$. This is a contradiction.

In *case* (2), the argument is similar to that in *case* (1). Note that if $|\rho| < 1$ then the subsequence $\{n_k\}$ should be made to go to $-\infty$, i.e. $0 > n_1 > n_2 > \ldots > -\infty$, to force $\rho^{n_k} \to \infty$. In *case* (3), the argument is the same as *case* (1) if $|\rho| = 1$; otherwise the argument is the same as for *case* (2). *Case* (4) can be treated similarly. □

A consequnce of Theorem 2.5.7 is that any rational cubic space curve that is affinely equivalent to the curve (2.45) can not be generated by any fourth order difference equation. An interesting open problem is, besides the polynomial cubic curves, which rational cubic curves can be generated by a fourth order difference equation.

2.6 Summary

In this chapter we have studied in detail properties of the difference method for drawing conic sections. It is shown that this method is more efficient as the matrix iteration method and generates the same point sequences on a conic than the latter. Compared with the forward differencing method, the difference method generates more evenly distributed point sequences on a conic. For instance, it can generate an equidistant point sequence on a circle; but this is not the case with the forward differencing method. Efficiencies of the difference method and the forward differencing method are comparable. On a conic in the plane, the former uses two multiplications and the latter uses two divisions to generate a point.

The numerical stability of the difference method is not discussed here since it has be addressed in [WaW89a] in the case of ellipses, and in [WaW89b] in the general case. We will just summarize some results here. When the conic is an ellipse, the difference method is more stable than the forward differencing method. It is shown in [WaW89a] that, when initial errors and roundoff errors are considered, the error of the difference method demonstrates a quadratic growth. On the other hand, the error growth of the forward differencing method is cubic. That is because, essentially, the forward differencing method uses a cubic difference equation, which has a cubic characteristic equation with a triple root 1. So, the error of the forward differencing method caused by initial errors alone has a quadratic growth, since in this case the error satisfies the same cubic difference equation used in forward differencing. Hence, when the contributions of roundoff errors at all points are taken into account, it can be shown that the total error growth is cubic. In the case of hyperbolas, the difference method has an exponential error growth. So in this case it is less stable than the forward differencing method.

We have discussed the problem of choosing the parameter k of the difference method to ensure that the inscribed polygon formed by the generated point sequence stays within a prescribed distance of a conic arc. We have considered how to use the difference method to draw conic sections given in the rational quadratic parametric representation or in the quadratic implicit form.

Finally, from a more theoretical point of view, we proved that the inscribed polygon generated by the difference method possesses a certain best approximation property. This property was found in [Smi71]. Here a new proof is given since the proof in [Smi71] is incorrect.

An important question about the difference method is whether or not it can be extended to generate other rational curves beside rational quadratic curves. We have shown that there exists no fourth order difference equation that can generate a point sequence on a certain class of rational cubic space curves. The case of rational plane cubic curves and other higher degree rational space curves is still open.

Chapter 3

Parametric Representations and Reparameterizations of Quadrics

In this chapter we study the reparameterization of rational surfaces and rational quadratic parameterizations of a quadric in E^3. First we will obtain a canonical form for a rational triangular Bézier surface under rational linear reparameterizations. Then we study some aspects of nonlinear reparameterizations in the setting of quadratic parameterizations of a quadric in E^3. The main results are a classification of all rational quadratic parameterizations of a quadric and their relationship. Finally, the above theory is applied to obtain a rational triangular Bézier representation of a triangular surface patch on a quadric whose boundaries are rational curve segments. The motivation for this study is mainly out of theoretical interests, but the triangular Bézier representation of a triangular patch on a quadric provides an alternate representation of the patches, and thus has applications in geometric modeling systems involving quadrics.

3.1 Preliminaries

A rational plane curve of degree n, in affine coordinates, is given by

$$P(t) = \left[\frac{x(t)}{w(t)}, \frac{y(t)}{w(t)}\right]^T,$$

where $x(t)$, $y(t)$ and $w(t)$ are polynomials with no common factor and $\max\{\deg(x(t)), \deg(y(t)), \deg(w(t))\} = n$. For a complete analysis, the parameter t takes values in C, the field of complex numbers, but in practice it is usually assumed that $x(t)$, $y(t)$ and $w(t)$ have real coefficients and t is in $\mathcal{R}$, the field of real numbers.

There are several other representations of rational curves, which may be convenient in certain circumstances. First we introduce the homogeneous representation of points in E^d, $d \leq 1$. A point in E^d can be represented by homogeneous coordinates $X = [x_1, \ldots, x_{d+1}]^T$, where at least one $x_i \neq 0$. Two $(d+1)$-tuples X_1, X_2 are defined to represent the same point in E^d if $X_1 = \rho X_2$ for some $\rho \neq 0$. If $x_{d+1} = 0$ then X is called a point at infinity with respect to E^d. A straight line passing through two distinct points X_0 and X_1 is denoted by X_0X_1, where X_0 or X_1 may be a point at infinity.

In homogeneous coordinates, a rational curve can be put in the form

$$P(t) = [x(t), y(t), w(t)]^T.$$

Or, if the homogeneous parameter pair (u, v) is used, we have

$$P(u,v) = [x(u,v), y(u,v), w(u,v)]^T, \tag{3.1}$$

where $x(u,v)$, $y(u,v)$ and $w(u,v)$ are homogeneous polynomials of degree n in u and v with no common factor.

A rational surface of degree n in E^3, in affine coordinates, is given by

$$P(s,t) = \left[\frac{x(s,t)}{w(s,t)}, \frac{y(s,t)}{w(s,t)}, \frac{z(s,t)}{w(s,t)}\right]^T,$$

where $x(s,t)$, $y(s,t)$, $z(s,t)$ and $w(s,t)$ are polynomials of degree n in s, t with no common factor and $\max\{\deg(x(s,t)), \deg(y(s,t)), \deg(z(s,t)), \deg(w(s,t))\} = n$. Similarly there is the homogeneous representation of a rational surface with homogeneous parameters (r, s, t), i.e.

$$P(r,s,t) = [x(r,s,t), y(r,s,t), z(r,s,t), w(r,s,t)]^T, \tag{3.2}$$

where $x(r,s,t)$, $y(r,s,t)$, $z(r,s,t)$ and $w(r,s,t)$ are homogeneous polynomials of degree n in r, s, t with no common factor. Other representations of a rational surface will be given when they are used.

For a rational parametric representation $P(s,t)$ of a surface S, if the correspondence of the parameters (s,t) with points on the surface is one-to-one, except perhaps at a finite number of curves on the surface, then the parameterization $P(s,t)$ is called *faithful*. A similar definition exists for rational curves and it is known that any rational curve possesses a faithful parameterization [Sed86]. But it is not known whether or not it is true of a rational surface.

An important concept in the study of rational surfaces is the *base point*. In the representation (3.2), (r_0, s_0, t_0) is defined to be a *base point* of the surface if $P(r_0, s_0, t_0) = 0$. Note that a rational curve has no base points, because for a curve, say in representation (3.1), if there was (u_0, v_0) such that $P(u_0, v_0) = 0$ then $x(u, v)$, $y(u, v)$ and $w(u, v)$ would have a common factor $(v_0 u - u_0 v)$.

The presence of base points in a rational surface plays an important role in the properties of the surface. It is shown [Chi90, p. 145] that a faithfully parameterized rational surface of degree n possessing b base points is an algebraic surface of degree $n^2 - b$. Also, when $P(r, s, t)$ is a rational surface of degree n, a rational curve of degree m in the parameter (r, s, t) domain passing through the base points of $P(r, s, t)$ p times is mapped by $P(r, s, t)$ to a rational curve of degree $mn - p$ [SeK52, p.321]. Therefore a general line in the parameter domain, i.e. a line not passing through any base point of $P(r, s, t)$, is mapped to a rational curve of degree n on the surface.

It is well known that a quadric surface $S \subset E^3$ can be represented as a rational quadratic surface $P(r, s, t)$ [Som51, p. 192]. The implication of the above results in the case of parametric representations of a quadric S is that any faithful rational quadratic parameterization of a quadric must have exactly two base points, say B_0 and B_1. Since an isolated base point is mapped to a line on S and all points on the line $B_0 B_1$ are mapped to a point X_0 on S, which is the intersection of the two lines corresponding to B_0 and B_1, the two base points B_0 and B_1 are mapped to the two generating lines of quadric S at X_0. As the quadric S is nondegenerate if and only if the two generating lines of S at X_0 (or any point on S) are distinct, the quadric S given by $P(r, s, t)$ is nondegenerate if and only if B_0 and B_1 are distinct. When the two base points coincide, S is called *properly degenerate*; in this case S is either a quadric cone or a conic cylinder. An example of a quadratic parameterization of a cylinder is given later in Ex. 3.3.1.

When $P(r, s, t)$ is a faithful rational parameterization of surface S, for a general point $X_0 \in S$, a formula that gives the corresponding parameter tuple (r, s, t) is called an *inversion formula* of the parameterization $P(r, s, t)$. Inversion formulas for rational surfaces are discussed in [ChR92].

3.2 Reparameterization

Let $P(s, t) = [x(s, t), y(s, t), z(s, t)]^T$ be a parametric surface in E^3. *Reparameterization* involves substituting for the parameters s and t in $P(s, t)$ by functions $s(u, v)$ and $t(u, v)$ so as to obtain a different parameterization $P(s(u, v), t(u, v))$ of the same surface in parameters

u and v. In the following we will consider only rational parametric surfaces $P(s,t)$, and assume that the substitutions $s(u,v)$, $t(u,v)$ are also rational functions. Sometimes the term *reparameterization* is also used to refer to the new parameterization $P(s(u,v),t(u,v))$. Reparameterization is similarly defined for rational curves where only one parameter, instead of a pair of them, is involved [Pat86].

Since a rational curve or surface has different parameterizations and the geometric features of a curve or surface does not depend on a particular parameterization, it is convenient to have a canonical parameterization of the rational curve or surface with respect to a certain class of its parameterizations. In the following we will discuss this canonical form under rational linear reparameterizations in the case of a rational triangular Bézier surface. Later we discuss reparameterizations that use substitutions of functions of degree > 1. Various aspects of these nonlinear rational reparameterizations are illustrated in the next section when we investigate rational quadratic parameterizations of a quadric in E^3. But before all this we first briefly review the canonical form of a rational curve [Pat86].

A general rational Bézier curve of degree n is given by

$$P(t) = \frac{\sum_{i=0}^{n} w_i P_i B_{i,n}(t)}{\sum_{i=0}^{n} w_i B_{i,n}(t)}, \quad 0 \le t \le 1, \tag{3.3}$$

where the w_i are called *weights*, the P_i are called *control points*, and the

$$B_{i,n}(t) = \frac{n!}{i!(n-i)!} t^i (1-t)^{n-i}, \quad i = 0, 1, \ldots, n,$$

are n-th degree Bernstein polynomials. When $w_0 > 0$ and $w_n > 0$, through a rational linear reparameterization given by

$$t = \frac{\alpha u}{\alpha u + \beta(1-u)},$$

where $\alpha = w_n^{-1/n}$ and $\beta = w_0^{-1/n}$, the same curve segment can be put in the form

$$P(t(u)) = \frac{\sum_{i=0}^{n} w_i' P_i B_{i,n}(u)}{\sum_{i=0}^{n} w_i' B_{i,n}(u)}, \quad 0 \le u \le 1,$$

where $w_i' = w_i \alpha^i \beta^{n-i}$, that is, the two corner weights $w_0' = w_n' = 1$. Note that when w_0 or w_n is zero, the denominator and the numerator have a common factor, meaning that the curve is of degree less than n. When $w_0 < 0$ and $w_n < 0$, changing the sign of the denominator and numerator in (3.3), we can set $w_0 > 0$ and $w_n > 0$. When $w_0 w_n < 0$, the denominator vanishes at a point in $[0,1]$, therefore the Bézier curve segment is not continuous. So our assumption that $w_0 > 0$ and $w_n > 0$ is not an essential restriction since in practice only continuous Bézier segments are of interest.

For a rational quadratic Bézier curve, the canonical form is

$$P(u) = \frac{P_0 B_{0,2}(u) + w_1 P_1 B_{1,2}(u) + P_2 B_{2,2}(u)}{B_{0,2}(u) + w_1 B_{1,2}(u) + B_{2,2}(u)}, \quad 0 \leq u \leq 1.$$

This representation will be used in Chapters 4 and 5.

3.2.1 Rational linear reparameterization

In this subsection we will derive the canonical form of a rational triangular Bézier surface under rational linear reparameterization, with the substitution functions mapping the parameter domain triangle onto itself.

Given three noncollinear points X_0, X_1, and X_2 in affine coordinates in the plane, any point X in the plane can be represented uniquely as $X = rX_0 + sX_1 + tX_2$, where $r+s+t = 1$. The tuple (r, s, t) is called the *barycentric coordinates* of X with respect to X_0, X_1, and X_2. A point X is inside the triangle $\Delta X_0 X_1 X_2$ iff $r \geq 0$, $s \geq 0$, and $t \geq 0$.

A rational triangular Bézier surface of degree n with parameters in barycentric coordinates is given by

$$P(r,s,t) = \frac{\sum_{i+j+k=n} w_{i,j,k} P_{i,j,k} B_{i,j,k,n}(r,s,t)}{\sum_{i+j+k=n} w_{i,j,k} B_{i,j,k,n}(r,s,t)}, \tag{3.4}$$

where $r+s+t = 1$, $r, s, t \geq 0$ and $B_{i,j,k,n} = \frac{n!}{i!j!k!} r^i s^j t^k$. The conditions $r, s, t \geq 0$ represents a triangle Δ in the parameter domain, explaining the name *triangular surface*. Δ will be called the *domain triangle*. Here we assume that the surface patch is continuous and $w_{n,0,0}$, $w_{0,n,0}$ and $w_{0,0,n}$ are all positive.

Now we seek a linear reparameterization of (3.4) so that under the new parameters the three corner weights $w_{n,0,0}$, $w_{0,n,0}$ and $w_{0,0,n}$ are 1. Suppose that the rational linear substitution functions for the required reparameterization are

$$\begin{aligned} r &= \frac{a_1 u + b_1 v + c_1 w}{a_4 u + b_4 v + c_4 w}, \\ s &= \frac{a_2 u + b_2 v + c_2 w}{a_4 u + b_4 v + c_4 w}, \\ t &= \frac{a_3 u + b_3 v + c_3 w}{a_4 u + b_4 v + c_4 w}. \end{aligned} \tag{3.5}$$

Since it is required that the three vertices $(u, v, w) = (1,0,0)$, $(0,1,0)$, and $(0,0,1)$ of triangle Δ be fixed by the substitution (3.6), we have

$$a_1 = a_4, \quad b_2 = b_4, \quad c_3 = c_4,$$

and

$$a_2 = a_3 = b_1 = b_3 = c_1 = c_2 = 0.$$

Let $a_4 = \alpha$, $b_4 = \beta$ and $c_4 = \gamma$. Then the substitution can be expressed as

$$\begin{aligned} r &= \frac{\alpha u}{\alpha u + \beta v + \gamma w}, \\ s &= \frac{\beta v}{\alpha u + \beta v + \gamma w}, \\ t &= \frac{\gamma w}{\alpha u + \beta v + \gamma w}. \end{aligned}$$

Here $\alpha, \beta, \gamma > 0$ can be assumed in order to map the interior of Δ onto itself. The tuple (α, β, γ), up to a common factor, characterizes a rational linear reparameterization that maps triangle Δ onto itself and keeps the three vertices fixed. Since $(u, v, w) = (1/3, 1/3, 1/3)$ is mapped to

$$(r, s, t) = \left(\frac{\alpha}{\alpha + \beta + \gamma}, \frac{\beta}{\alpha + \beta + \gamma}, \frac{\gamma}{\alpha + \beta + \gamma}\right),$$

the image of the center of Δ in the r, s, t plane also characterizes the rational linear reparameterization indicated above.

Applying the above linear reparameterization to the rational surface (3.4) we have another parameterization of the same surface

$$\begin{aligned} \bar{P}(u, v, w) &= \frac{\sum_{i+j+k=n} w_{i,j,k}\alpha^i\beta^j\gamma^k P_{i,j,k}B_{i,j,k,n}(u, v, w)}{\sum_{i+j+k=n} w_{i,j,k}\alpha^i\beta^j\gamma^k B_{i,j,k,n}(u, v, w)} \\ &\equiv \frac{\sum_{i+j+k=n} w'_{i,j,k}P_{i,j,k}B_{i,j,k,n}(u, v, w)}{\sum_{i+j+k=n} w'_{i,j,k}B_{i,j,k,n}(u, v, w)}. \end{aligned} \tag{3.6}$$

Let $\alpha = (w_{n,0,0})^{-1/n}$, $\beta = (w_{0,n,0})^{-1/n}$ and $\gamma = (w_{0,0,n})^{-1/n}$. Then the three corner weights $w'_{n,0,0}$, $w'_{0,n,0}$ and $w'_{0,0,n}$ in (3.6) are 1. In the case of rational quadratic triangular Bézier patches, i.e. $n = 2$, we can reduce the six weights in a general parameterization to three weights $w_{1,1,0}, w_{0,1,1}$ and $w_{1,0,1}$ using a linear reparameterization.

In the above it was assumed that the three corner weights are positive. When some corner weight is zero, the parameter values at that corner are a base point of the rational surface, i.e. parameter values that make all the numerators and the denominator in the right hand side of (3.4) vanish. Geometrically, in this case, if it is the only base point of the surface in Δ, the parameter domain triangle is no longer mapped to a triangular surface patch, but to a four-sided patch. The general case and the application of this singularity in surface modeling are discussed in [War90]. The case that two of $w_{n,0,0}$, $w_{0,n,0}$ and $w_{0,0,n}$ are of opposite signs is excluded by the assumption that the surface patch is continuous.

3.2.2 General reparameterization

A reparameterization of a rational surface can be carried out with rational functions of degree higher than one. And in this case, the new parameterization may still be faithful, i.e. there is a one-to-one correspondence between points on the surface and the points in the parameter plane, except perhaps on a finite number of curves. This is in contrast with the case of rational curves, where for a faithfully parameterized rational curve $P(t)$, if the substitution $t = t(u)$ is of degree $n > 1$, the correspondence between the parameter values u and points on the curve $P(t(u))$ is n-to-one; therefore the curve is unfaithfully parameterized [JoW90].

That a nonlinear rational reparameterization of a faithfully parameterized surface can still be faithful follows from the fact that there exist rational transformations of any degree of the parameter plane which are one-to-one, except perhaps on a finite number of curves. But on the straight line a one-to-one rational transformation must be rational linear. The theory of general one-to-one rational transformations of the plane, called *Cremona transformations*, have been extensively studied in algebraic geometry [Sal34, p. 314, SeK52, p. 230].

Let

$$s = \frac{f_1(u,v)}{g(u,v)}, \quad t = \frac{f_2(u,v)}{g(u,v)}$$

be a rational transformation of degree n in the parameter domain. Then it is known that this transformation is a Cremona transformation iff it has $n^2 - 1$ base points [Sal34, p. 314], i.e. there are $n^2 - 1$ points (u, v) where the polynomials f_1, f_2 and g vanish simultaneously. The argument is basically as follows. We just need to show that corresponding to a general point (s_0, t_0) there is only one associated point (u, v) when the mapping has $n^2 - 1$ base points. Clearly any point (u, v) that is mapped to (s_0, t_0) satisfies

$$s_0 g(u,v) - f_1(u,v) = 0, \quad t_0 g(u,v) - f_2(u,v) = 0.$$

Therefore the point (u, v) is an intersection of these two curves, which have n^2 intersections in all. Since $n^2 - 1$ of them are base points, which are fixed, there is only one free intersection, denoted by (u_0, v_0), which corresponds to (s_0, t_0). This argument can be made complete by using homogeneous parameters, so that points at infinity are also accounted for.

From the above discussion it follows that a faithful parameterization of a rational surface does not necessarily have minimum degree. In fact, the existence of a nonlinear Cremona transformation on a plane is a nonlinear rational faithful parameterization of the plane, which, as a surface, has a faithful linear parameterization. A nontrivial example is the following faithful rational cubic parameterization, in affine coordinates, for the cylinder

$x^2 + y^2 = 1$,

$$[x, y, z]^T = \left[\frac{1-u^2}{1+u^2}, \frac{2u}{1+u^2}, v\right]^T.$$

In the next section we will see that this cylinder has a faithful quadratic parameterization.

There are many unanswered questions about the parameterizations of a rational surface. For instance, does a rational surface always have a faithful parameterization? Are any two parameterizations of minimum degree of a rational surface related by a rational linear reparameterization? The answer to the first question is unclear. We will see in the following sections that two faithful parameterizations of minimum degree of a rational surface are not necessarily related by a rational linear reparameterization.

3.3 Rational parameterizations of quadrics

In this section we study rational quadratic parameterizations of a nondegenerate or properly degenerate quadric S given by $X^T AX = 0$ in E^3, i.e. rank(A)= 4 or 3, where A is a 4×4 symmetric matrix and X is in homogeneous coordinates. We also consider how different quadratic parameterizations of S are related by a reparameterization. It is known that a quadric can be represented as a rational quadratic surface [Som51, p. 192]. Here we shall study the structure of all quadratic parameterizations of a quadric S. It will be shown that (1) associated with any point on a quadric S in E^3 there is a family of quadratic parameterizations of S; (2) any rational quadratic parameterization of S is associated with a point on S; (3) any quadratic parameterization of S has a rational linear inversion formula, which can be interpreted as a projection from the associated point of the parameterization; (4) two quadratic parameterizations of S associated with the same point are related by a rational linear reparameterization; (5) two quadratic parameterizations of S with different associated points are related by a rational quadratic reparameterization.

3.3.1 Parameterizations by projection

Given a quadric surface S : $X^T AX = 0$, we first derive one of its quadratic parameterizations. The approach we use is the stereographical projection of a quadric surface [Som51, p. 192], i.e. choose a point X_0 on S and project other points on S onto a plane not passing through X_0. Point X_0 is called the *center of projection* (COP) of the parameterization. Although this parameterization is well known, here we are interested in the relationship of all these parameterizations.

Let $X_0 \in S$. Then all lines passing through X_0 are parameterized by $P_T(u,v) = uX_0 + vT$, where T is a variable point on a plane $B_0^T X = 0$ not passing through X_0, i.e. $B_0^T X_0 \neq 0$, and u, v are homogeneous parameters of straight line X_0T. The line $P_T(u,v)$, for fixed T, has two intersections with S, one of which is X_0. To find the other, substituting $P_T(u,v)$ in $X^T AX = 0$, we have

$$u^2 X_0^T AX_0 + 2uvX_0^T AT + v^2 T^T AT = 0.$$

Since $X_0^T AX_0 = 0$, the other intersection corresponds to the parameters (u_T, v_T)=$(T^T AT, -2X_0^T AT)$. So a rational parameterization of S, in homogeneous coordinates, is

$$P(T) \equiv P_T(u_T, v_T) = (T^T AT)X_0 - 2(X_0^T AT)T, \tag{3.7}$$

where T, restricted to the plane $B_0^T X = 0$, stands for the rational parameters of the surface.

When $T^T AT = T^T AX_0 = 0$, $(u_T, v_T) = (0,0)$, which are not well defined homogeneous parameter values. Let T_0 and T_1 be the two solutions of T solved from $T^T AT = T^T AX_0 = 0$. Then T_0 and T_1 are the two base points of $P(T)$. Since $X_0^T AX = 0$ is the tangent plane of S at X_0, the straight lines X_0T_0 and X_0T_1 are the two generating lines of S at X_0. So the two base points T_0 and T_1 of $P(T)$ are mapped to the two generating lines at X_0 in a general sense.

To obtain the inversion formula for (3.7), consider a general point $X \in S$ which is not X_0. Let the corresponding parameter be $T = \lambda_1 X_0 + \mu_1 X$. Then from

$$0 = B_0^T T = \lambda_1 B_0^T X_0 + \mu_1 B_0^T X,$$

we have

$$T = (B_0^T X)X_0 - (B_0^T X_0)X, \tag{3.8}$$

which is a rational linear inversion formula of (3.7). Let (r,s,t) be a projective coordinate system on the plane $B_0^T X = 0$. Since T is a point on this plane, there exists a 4×4 nonsingular matrix M such that $T = M[r,s,t,0]^T$; in fact, the first three columns of M represent the three reference points of the projective system (r,s,t) of the plane $B_0^T X = 0$, and the fourth column is any point not on the plane. Now r,s,t can be regarded as homogeneous parameters of the representation (3.7). An inversion formula equivalent to (3.8) is

$$[r,s,t,0]^T = M^{-1}[(B_0^T X)X_0 - (B_0^T X_0)X]. \tag{3.9}$$

Example 3.3.1: We illustrate the above results in the case where S is the cylinder S: $X^T AX = 0$, with $A = \text{diag}[1,1,0,-1]$ and $X = [x,y,z,w]^T$ being homogeneous coordinates. Let the COP $X_0 = [1,0,0,1]^T$, and the parameter plane be $B_0^T X = 0$ where

$B_0 = [1, 0, 0, 0]^T$. Let $T = [0, r, s, t]^T$, where r, s, t are homogeneous parameters. Then, by (3.7), a quadratic parameterization of S is

$$[x, y, z, w]^T = [r^2 - t^2, 2rt, 2st, r^2 + t^2]^T. \tag{3.10}$$

It is easy to check that $(r, s, t) = (0, 1, 0)$ is a double base point of the above parameterization. For an inversion formula we see that

$$T = [0, r, s, t]^T = M[r, s, t, 0]^T,$$

where

$$M = \begin{bmatrix} 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix}.$$

Then by (3.9) we have

$$[r, s, t, 0]^T = [-y, -z, x - w, 0]^T.$$

So an inversion formula for (3.10) is

$$(r, s, t) = (y, z, w - x). \tag{3.11}$$

Substituting in the parameterization (3.10), we have

$$(y, z, w - x) = 2t(r, s, t).$$

So (3.11) is a correct inversion formula, which is invalid when $y = z = w - x = 0$, i.e. when $[x, y, z, w]^T = [1, 0, 0, 1]^T$, which is the COP. Obviously, all points $[1, 0, z, 1]^T$, $z \neq 0$, on the double generating lines of S at COP $[1, 0, 0, 1]^T$ have the same parameter tuple $(0, 1, 0)$. □

Now we consider how two quadratic parameterizations of the same quadric as obtained above are related through reparameterization. Suppose that we have two parameterizations of S,

$$P(T) = (T^T AT)X_0 - 2(X_0^T AT)T, \tag{3.12}$$

and

$$Q(U) = (U^T AU)X_1 - 2(X_1^T AU)U, \tag{3.13}$$

both of which are obtained using the stereographic projection, where T and U are restricted to the planes $B_0^T X = 0$ and $B_1^T X = 0$, respectively. Furthermore we assume that $B_0^T X_0 \neq 0$ and $B_1^T X_1 \neq 0$. Now we shall find a reparameterization relating the parameters T and U. By (3.8), the parameter T of the point $Q(U)$ generated through (3.12) is

$$T = [B_0^T Q(U)]X_0 - (B_0^T X_0)Q(U),$$

or, substituting in $Q(U)$,

$$T = (U^T AU)[(B_0^T X_1)X_0 - (B_0^T X_0)X_1] - 2(X_1^T AU)[(B_0^T U)X_0 - (B_0^T X_0)U]. \tag{3.14}$$

This is the reparameterization relating U and T. Conversely, we have

$$U = (T^T AT)[(B_1^T X_0)X_1 - (B_1^T X_1)X_0] - 2(X_0^T AT)[(B_1^T T)X_1 - (B_1^T X_1)T]. \tag{3.15}$$

In general, the reparameterization (3.14) is quadratic. Obviously, it becomes linear if $X_0 = X_1$, i.e. $P(T)$ and $Q(U)$ have the same COP, because in this case

$$T = -2(X_1^T AU)[(B_0^T U)X_0 - (B_0^T X_0)U],$$

and the common factor $-2(X_1^T AU)$ can be omitted. Hence (3.14) becomes linear if $P(T)$ and $Q(U)$ have the same COP. We will show later that (3.14) is quadratic if $P(T)$ and $Q(U)$ have different COP's. Therefore two minimum degree parameterizations of a quadric surface are not necessarily related by a rational linear reparameterization.

Example 3.3.2: Here we consider the following two parameterizations of sphere S: $X^T AX = 0$, with $A = \text{diag}[1, 1, 1, -1]$. Let the two COP's be $X_0 = [0, 0, 1, 1]^T$ and $X_1 = [0, 0, -1, 1]^T$. Let the two parameter planes be $B_0^T X = 0$ and $B_1^T X = 0$, where $B_0 = B_1 = [0, 0, 1, 0]^T$, i.e. the two planes are the same. Let $T = [r, s, 0, t]^T$ and $U = [u, v, 0, q]^T$ be homogeneous parameters, restricted in the planes $B_0^T X = 0$ and $B_1^T X = 0$, respectively. By (3.7) we obtain two parameterizations of S

$$P(T) = [2rt, 2st, r^2 + s^2 - t^2, r^2 + s^2 + t^2]^T$$

and

$$Q(U) = [2uq, 2vq, -u^2 - v^2 + q^2, u^2 + v^2 + q^2]^T.$$

By (3.14),

$$T(U) = [-2uq, -2vq, 0, -2(u^2 + v^2)]^T.$$

So the quadratic reparameterization from U to T is

$$(r, s, t) = (uq, vq, u^2 + v^2),$$

which is a Cremona transformation as it has three base points $(u, v, q) = (0, 0, 1)$ and $(1, \pm i, 0)$. □

The reparameterizations (3.14) and (3.15) are Cremona transformations that are inverse to each other. According to the above discussion, they each should have three base points when X_0 and X_1 are different points. For (3.14) it is easy to verify that its three base points are the two intersections of $U^T AU = 0$ with $X_1^T AU = 0$ (and $B_1^T U = 0$), denoted by U_0 and U_1, respectively, and $U_2 = (B_1^T X_0)X_1 - (B_1^T X_1)X_0$, i.e. these three points make $T(U)$ vanish. The three base points of (3.15) are the two intersections of $T^T AT = 0$ with $X_0^T AT = 0$ (and $B_0^T T = 0$), denoted by T_0 and T_1, respectively, and $T_2 = (B_0^T X_1)X_0 - (B_0^T X_0)X_1$.

It is interesting to note that the composition of quadratic parameterization (3.12) and quadratic reparameterization (3.14) is the quadratic parameterization (3.13), instead of a quartic one as expected in general. We shall briefly examine the role played by the base points of (3.12) and those of (3.14) in this regard.

The three base points of (3.14), i.e. U_0, U_1 and $U_2 = X_1 - X_0$, determine three lines U_0U_1, U_0U_2 and U_1U_2. These three lines are mapped onto three points in the parameter T domain through $T(U)$ (3.14). From [SeK52, p.231], we know that the three base points of $T(U)$ are mapped onto three lines in the T plane, and these three lines intersect pairwise at the three base points of $U(T)$ (3.15), which is the inverse of $T(U)$. Therefore two of these three points are the intersections of $T^T AT = 0$ and $X_0^T AT = 0$, which have been shown to be the base points of $P(T)$. Now let us consider a general line l in parameter U domain. First, l is mapped by $T(U)$ to a rational quadratic curve l' on S. Since l intersects the three lines U_0U_1, U_0U_2 and U_1U_2, l' passes through the two base points T_0 and T_1 of $P(T)$, therefore l' is mapped by $P(T)$ to a rational quadratic curve on S. Thus a general line in U domain is mapped by $Q(U) = P(T(U))$ onto a rational quadratic curve on S. Hence $Q(U)$ is a rational quadratic parameterization of S.

3.3.2 General parameterizations

So far we have considered only the rational quadratic parameterizations that are obtained explicitly using stereographic projection. Now we ask whether any rational quadratic parameterization of S can be obtained in this way. By answering this question, we will gain a better understanding of quadratic parameterizations of a quadric S and their mutual relationships.

There exist unfaithful rational quadratic parameterizations for some quadric surface. For instance, the cone $x^2 + y^2 = z^2$ has the unfaithful parameterization $x = r^2 - s^2$, $y = 2rs$, $z = r^2 + s^2$, and $w = t^2$. This parameterization maps two parameter points into a point on the cone, and has no base points. In the following discussion we will consider only faithful parameterizations.

Theorem 3.3.1: *Let S be a quadric surface in E^3 and $P(r,s,t)$ be a faithful rational quadratic parameterization of S with homogeneous parameters r,s,t. Then there exists a* *rational linear inversion formula of $P(r,s,t)$.*

Proof: First suppose that S is nondegenerate. Then $P(r,s,t)$ has two distinct base points B_0 and B_1. Select a point B_2 which is not on the line B_0B_1. Then B_0, B_1 and B_2 determine three lines l_0: B_1B_2 , l_1 : B_0B_2 and l_2 : B_0B_1. We will use $l_i(r,s,t)=0$ to denote the linear equation of the line l_i, $i=0,1,2$.

Let $P(r,s,t)=[x(r,s,t),y(r,s,t),z(r,s,t),w(r,s,t)]^T$. Then $x(r,s,t),y(r,s,t),z(r,s,t)$, $w(r,s,t)$ are linearly independent. For if

$$c_1x(r,s,t)+c_2y(r,s,t)+c_3z(r,s,t)+c_4w(r,s,t)=0$$

with the c_i being not all zero, then S is contained in the plane

$$c_1x+c_2y+c_3z+c_4w=0,$$

which is impossible.

All conics passing through the two base points B_0 and B_1 form a projective space $\mathcal{L}$ of dimension 3, where the coefficients of the quadratic equation of a conic are regarded as homogeneous coordinates for the conic. Since $x(r,s,t),y(r,s,t),z(r,s,t)$ and $w(r,s,t)\in\mathcal{L}$ are linearly independent, their linear combinations span the space $\mathcal{L}$. Now consider three conics in $\mathcal{L}$: $l_0(r,s,t)l_2(r,s,t)=0$, $l_1(r,s,t)l_2(r,s,t)=0$, and $l_2^2(r,s,t)=0$. Then we have

$$\begin{aligned} l_0(r,s,t)l_2(r,s,t) &= a_{01}x(r,s,t)+a_{02}y(r,s,t)+a_{03}z(r,s,t)+a_{04}w(r,s,t),\\ l_1(r,s,t)l_2(r,s,t) &= a_{11}x(r,s,t)+a_{12}y(r,s,t)+a_{13}z(r,s,t)+a_{14}w(r,s,t),\\ l_2^2(r,s,t) &= a_{21}x(r,s,t)+a_{22}y(r,s,t)+a_{23}z(r,s,t)+a_{24}w(r,s,t), \end{aligned} \tag{3.16}$$

for some constants a_{ij}. Thus for a point $[x,y,z,w]^T=[x(r,s,t),y(r,s,t),z(r,s,t),$ $w(r,s,t)]^T$ on S,

$$\begin{aligned} l_0(r,s,t)l_2(r,s,t) &= a_{01}x+a_{02}y+a_{03}z+a_{04}w \equiv L_1(x,y,z,w),\\ l_1(r,s,t)l_2(r,s,t) &= a_{11}x+a_{12}y+a_{13}z+a_{14}w \equiv L_2(x,y,z,w),\\ l_2^2(r,s,t) &= a_{21}x+a_{22}y+a_{23}z+a_{24}w \equiv L_3(x,y,z,w). \end{aligned} \tag{3.17}$$

As $l_0(r,s,t)$, $l_1(r,s,t)$ and $l_2(r,s,t)$ are linear in homogeneous parameters r, s and t, after dividing by $l_2(r,s,t)$, we can solve for an inversion formula from the above system in the form,

$$\begin{aligned} r &= b_{11}x+b_{12}y+b_{13}z+b_{14}w \equiv A_1(x,y,z,w)\\ s &= b_{21}x+b_{22}y+b_{23}z+b_{24}w \equiv A_2(x,y,z,w)\\ t &= b_{31}x+b_{32}y+b_{33}z+b_{34}w \equiv A_3(x,y,z,w), \end{aligned} \tag{3.18}$$

for some constants b_{ij}, $i = 1,2,3$, $j = 1,2,3,4$.

When S is properly degenerate, the two base points B_0 and B_1 coincide. So the four conics $x(r,s,t) = 0$, $y(r,s,t) = 0$, $z(r,s,t) = 0$ and $w(r,s,t) = 0$ in the parameter domain have a common tangent at $B_0 = B_1$. This common tangent can then be used as the line $l_2(r,s,t) = 0$ as in the above argument, and the lines l_0 and l_1 are chosen so that they do not intersect in a point on l_2, but otherwise arbitrary. The remaining proof follows identically. □

It is worthwhile to point out the singularity with regard to inversion formula (3.18), i.e. the cases where (3.18) does not give a one-to-one correspondence. First we note that the choices of l_0 and l_1 are arbitrary, except that they are distinct and do not coincide with l_2: B_0B_1, with l_2 being defined appropriately when $B_0 = B_1$. The inversion formula (3.18) determines two pencils of planes by r, s and t, which are

$$\begin{aligned} rA_3(x,y,z,w) - tA_1(x,y,z,w) = 0, \\ sA_3(x,y,z,w) - tA_2(x,y,z,w) = 0. \end{aligned} \tag{3.19}$$

Since the line l_2 passes through the two base points B_0 and B_1, its image under $P(r,s,t)$ is a single point on S. Let $X_0 \in S$ be this point. Then from (3.16) and (3.17), we have $L_1(X_0) = L_2(X_0) = L_3(X_0) = 0$. Therefore the $A_i(X_0) = 0$ since the $A_i(x,y,z,w)$ are linear combinations of the $L_i(x,y,z,w)$, $i = 1,2,3$. Hence no definite parameters r,s,t can be computed for from (3.18) to correspond to X_0. Geometrically, the axes of the two pencils in (3.19) both pass through X_0. From Ex. 3.3.1 we see that the COP X_0 of parameterization (3.10) is given by the solution of $(y, z, w - x) = (0,0,0)$, i.e. $X_0 = [1,0,0,1]^T$.

Another singularity is for points on the two generating lines of S passing through X_0. Let g_0 be one of the two. Then there is a plane p_0: $r_0A_3 - t_0A_1 = 0$ and a plane p_1: $s_0A_3 - t_0A_2 = 0$ belonging to the two pencils, respectively, such that p_0 and p_1 both contain the generating line g_0. Therefore all points on g_0, except X_0, have the same parameter value (r_0, s_0, t_0).

The correspondence of other points on S with parameter tuples (r,s,t) within a constant factor is one-to-one, and for such a point on S a unique parameter triple, up to a proportional constant, can be computed from (3.18). Since the axes of the two pencils in (3.19) pass through X_0, for any fixed parameters r_1, s_1 and t_1 two corresponding planes of the two pencils intersect in a line passing through X_0. This line intersects the quadric S in another point X_1 besides X_0, so X_1 corresponds to the parameters (r_1, s_1, t_1). Hence $P(r,s,t)$ can be obtained through a stereographic projection, and the associated point of a quadratic parameterization of S can be identified with its COP. Note that here the plane of projection can be any plane not passing through X_0, and (r, s, t) is an appropriate projective coordinate system in that plane.

Now we consider the reparameterization that relates two different faithful quadratic parameterizations of a quadric in E^3. Let

$$P(r,s,t) = [x(r,s,t), y(r,s,t), z(r,s,t), w(r,s,t)]^T,$$
$$Q(\bar{r},\bar{s},\bar{t}) = [\bar{x}(\bar{r},\bar{s},\bar{t}), \bar{y}(\bar{r},\bar{s},\bar{t}), \bar{z}(\bar{r},\bar{s},\bar{t}), \bar{w}(\bar{r},\bar{s},\bar{t})]^T$$

be two faithful rational quadratic parameterizations of a quadric $S \subset E^3$. Let the associated points of $P(r,s,t)$ and $Q(\bar{r},\bar{s},\bar{t})$ be $X_0, X_1 \in S$, respectively. According to Theorem 3.3.1, we have an inversion formula of $P(r,s,t)$

$$\begin{aligned} r &= A_1(x,y,z,w), \\ s &= A_2(x,y,z,w), \\ t &= A_3(x,y,z,w), \end{aligned}$$

where A_1, A_2, A_3 are linear. Substituting in the components of $Q(\bar{r},\bar{s},\bar{t})$ for x, y, z, and w in the above inversion formula, we have the reparameterization that relates $(\bar{r},\bar{s},\bar{t})$ and (r,s,t)

$$\begin{aligned} r &= A_1(\bar{x}(\bar{r},\bar{s},\bar{t}), \bar{y}(\bar{r},\bar{s},\bar{t}), \bar{z}(\bar{r},\bar{s},\bar{t}), \bar{w}(\bar{r},\bar{s},\bar{t})), \\ s &= A_2(\bar{x}(\bar{r},\bar{s},\bar{t}), \bar{y}(\bar{r},\bar{s},\bar{t}), \bar{z}(\bar{r},\bar{s},\bar{t}), \bar{w}(\bar{r},\bar{s},\bar{t})), \\ t &= A_3(\bar{x}(\bar{r},\bar{s},\bar{t}), \bar{y}(\bar{r},\bar{s},\bar{t}), \bar{z}(\bar{r},\bar{s},\bar{t}), \bar{w}(\bar{r},\bar{s},\bar{t})). \end{aligned} \tag{3.20}$$

They are, in general, rational quadratic functions. Since, by symmetry, we can express $\bar{r}$, $\bar{s}$, and $\bar{t}$ as rational quadratic functions in r, s and t, (3.20) is a Cremona transformation.

Theorem 3.3.2: *The reparameterization expressed by (3.20) is rational linear if and only if $X_0 = X_1$, i.e they have the same COP.*

Since by using inversion formulas we have shown that every faithful rational quadratic parameterization of a quadric can be obtained by a projection through its COP, the sufficiency of Theorem 3.3.2 has been in fact proved in Section 3.3.1. Here we will give another geometric proof for it.

PROOF: Suppose that X_0 and X_1 are the same point. To show that (3.20) is in fact rational linear, we just have to prove that a general line in $(\bar{r},\bar{s},\bar{t})$ domain is mapped by (3.20) onto a line in (r,s,t) domain.

Let l be a general line in $(\bar{r},\bar{s},\bar{t})$ domain and let B_0 and B_1 be the two base points of $Q(\bar{r},\bar{s},\bar{t})$. Since l intersects the line B_0B_1, which is mapped to $X_1 \in S$, l is mapped to a conic on S passing through X_1; when $B_0 = B_1$ the line B_0B_1 is defined appropriately as in the proof of Theorem 3.3.1. Conversely, any conic on S passing through X_1 is the image of a line in $(\bar{r}, \bar{s}, \bar{t})$ domain; that is because, as indicated above, any quadratic parameterization is obtained by stereographic projection. Therefore there is a one-to-one correspondence between all general lines in $(\bar{r}, \bar{s}, \bar{t})$ domain and all conics on S passing through X_1. As the same correspondence also holds for the parameterization $P(r,s,t)$ of S and $X_0 = X_1$, we conclude that a general line in $(\bar{r}, \bar{s}, \bar{t})$ domain corresponds to a line in (r,s,t) domain through the mapping $P^{-1}Q$. Hence the reparameterization expressed by (3.20) is rational linear.

The necessity can be proved similarly. In the above argument, when X_0 and X_1 are different, a general conic on S passing through X_1 projects from X_0 to a conic in the (r,s,t) plane, and thus it corresponds to a conic in the (r,s,t) plane. So a general line in the $(\bar{r}, \bar{s}, \bar{t})$ plane is mapped by $P^{-1}Q$ onto a conic in the (r,s,t) plane. Hence (3.20) is quadratic. $\square$

Let $P(r,s,t)$ be a quadratic parameterization of a quadric S obtained by projection from a point $X_0 \in S$. The two generating lines of S at X_0 then correspond to the two base points of $P(r,s,t)$. So the two base points are real if and only if the two generating lines of S at X_0, or equivalently at any other point on S, are real. For example, the following parameterization of the unit sphere

$$P(r,s,t) = [2rt,\ 2st,\ t^2 - r^2 - s^2,\ t^2 + r^2 + s^2]^T,$$

has imaginary base points $(1, \pm i, 0)$.

3.4 Bézier representation of a triangular patch on a quadric

In this section we consider the following problem: given a triangular patch on a quadric $S \subset E^3$ whose boundaries are rational curve segments, find a rational triangular Bézier representation of the patch. This problem has applications in surface modeling using quadrics [Dah89]. Several works address problems similar to the above. In [FPW88] an octant of a sphere is represented as a rational quartic Bézier surface. In [Sed85] triangular patches on a quadric that can be represented as rational quadratic Bézier surfaces are considered. In [JoW90] triangular patches on quadrics with conic arcs as boundaries are represented as rational quartic Bézier surfaces.

The main results of this section are: (1) using stereographic projection and reparameterization, a triangular patch T on a quadric S whose boundaries are rational curve segments of degree at most n can be represented as a Bézier surface of degree at most $2n$; (2) the

Bézier representation of $\mathcal{T}$ is quadratic if and only if there exist three planes each containing a boundary segment of $\mathcal{T}$ such that the three planes intersect at a point on S. The second result is proved in [LoW90]. The necessity part of result (2) is also proved in [Sed85]. But our approach is different in that we do not transform the quadric in question into the surface of a quadratic function as done in [LoW90].

As an application of the above results, we will show that any rational triangular Bézier representation for an octant of a sphere is of degree at least 4. Therefore the representation given in [FPW88] is of minimum degree.

3.4.1 General triangular patches

Let $\mathcal{T}$ be a triangular patch on a quadric S bounded by the rational curve segments $q_i(t)$ of degree at most n, $i = 0, 1, 2$, with at least one of the $q_i(t)$ of exact degree n. By a triangular patch we mean that each side of the patch does not intersect itself, and two sides of the patch intersect exactly once. Assume that there exists a point $X_0 \in S$ outside $\mathcal{T}$ so that none of the generating lines of S at X_0 intersects $\mathcal{T}$. Note that when S has only imaginary generating lines, like a sphere, any point $X_0 \in S$ outside $\mathcal{T}$ satisfies this assumption. Let p be a plane that does not pass through X_0 and does not intersect $\mathcal{T}$. Let F be the projection from E^3 to the plane p with center of projection on X_0, and let $\mathcal{T}' = F(\mathcal{T})$. From the assumption that none of the generating lines at X_0 intersects $\mathcal{T}$, and the fact that any line passing through X_0 intersects S in exactly one other point unless it lies entirely on S, the mapping $F|_{\mathcal{T}}$, from $\mathcal{T}$ to $\mathcal{T}'$ is one-to-one, where $F|_{\mathcal{T}}$ is the restriction of F to $\mathcal{T}$. Hence $\mathcal{T}'$ is also a triangular patch.

Let the boundary segments of $\mathcal{T}'$ be $\bar{q}_i(t)$, $i = 0, 1, 2$. Since the projection F is rational linear, $\bar{q}_i(t)$ is of the same degree as $q_i(t)$, $i = 0, 1, 2$. Now it is straightforward to construct a rational mapping $R(r, s, t)$ of degree n in Bézier form from a domain triangle Δ to $\mathcal{T}'$. We just have to treat $\mathcal{T}'$ as a planar Bézier surface patch of degree n, with the interior control points chosen appropriately. Let $P(r, s, t)$ be the quadratic parameterization of S obtained by projection with COP at X_0, which is the inversion of the mapping $F|_S$, the restriction of F to quadric S. Then the triangular patch $\mathcal{T}$ can be readily represented as the triangular Bézier surface $P(R(r, s, t))$ of degree $2n$.

Although it is easy to obtain a rational mapping from the domain triangle Δ to the planar patch $\mathcal{T}'$, in general it is unknown how to make this mapping one-to-one. Conditions that guarantee that this mapping is one-to-one in the case that the boundaries of $\mathcal{T}$ are conic arcs are given in [JoW90].

With this construction, an octant of a sphere is represented by a rational quartic Bézier surface in [FPW88].

3.4.2 Quadratic Bézier surfaces on a quadric

In this subsection we consider a special triangular patch $\mathcal{T}$ on a quadric S.

Theorem 3.4.1: *A triangular patch $\mathcal{T}$ on quadric $S \subset E^3$ can be represented as a faithful quadratic Bézier surface if and only if there exist three planes each containing a boundary segment $q_i(t)$ of $\mathcal{T}$, $i = 0, 1, 2$, such that the three planes intersect at a point on S.*

The necessity of Theorem 3.4.1 is proved in [Sed85]. The sufficiency is proved in [JoW90]. This result is also proved in [LoW90]. Here we give a proof without transforming the quadric S into the surface of a quadratic function as done in [LoW90].

PROOF: Sufficiency: Suppose that $\mathcal{T}$ is a triangular patch on S with the three boundary segments respectively contained in three planes which intersect at a point $X_0 \in S$. Select a plane p that does not pass through X_0 and does not intersect $\mathcal{T}$. Then the image $\mathcal{T}'$ of $\mathcal{T}$ on plane p under the projection through X_0 is a triangle. Now parameterizing S using the projection with COP being at X_0, we can obtain a quadratic Bézier representation $P(r, s, t)$ of $\mathcal{T}$ with $\mathcal{T}'$ as the parameter domain triangle.

Necessity: See [Sed85]. □

Theorem 3.4.2: *An octant of a sphere can be represented by a rational quartic triangular Bézier patch; and this representation is of minimum degree.*

PROOF: A quartic representation has been given in [FPW88]. We just have to prove that the part $\mathcal{T}$ of the unit sphere S^2 in the first octant of the coordinate system, i.e. where $x, y, z \geq 0$, can not be represented by a rational quadratic or cubic triangular Bézier patch.

Let $P(r, s, t)$ be a rational triangular Bézier representation of the triangular patch $\mathcal{T}$. Obviously, $P(r, s, t)$ can not be a faithful quadratic parameterization. This follows directly from Theorem 3.4.1, since the three planes determined by the three boundary circular arcs of $\mathcal{T}$ intersect in the center of the sphere, which is not on the sphere.

Secondly, we need to show that $P(r, s, t)$ is not an unfaithful quadratic parameterization. Suppose that $P(r, s, t)$ is unfaithful. Then $P(r, s, t)$ has at most one base point. Choose a general real point X on $\mathcal{T}$ so that there are two distinct parameter points U and V in the plane (r,s,t) corresponding to the point $X \in \mathcal{T}$ through $P(r,s,t)$. Note that, since $P(r,s,t)$ is quadratic, there are at most two distinct parameter points corresponding to the general point X. Obviously, one of U and V is real, since X corresponds to at least one real parameter point. Let U be real. Then we claim that V is also real. From $P(V) = X$, we have $P(\bar{V}) = \bar{X} = X$, where $\bar{V}$ is the complex conjugate point of V. If V is not real, then V, $\bar{V}$, and U are three distinct parameter points corresponding to X through $P(r,s,t)$. This contradicts the assumption that $P(r,s,t)$ is quadratic.

Let L be the straight line passing through the points U and V. Then L is a real line. The image $P(L)$ of L on S^2 is a real straight line if L passes through one base point of $P(r,s,t)$; and $P(L)$ is a real conic on S^2 if L passes through no base point of $P(r,s,t)$. The former case can be excluded since there is no real line on S^2. In the latter case, we can exclude the possibility that $P(L)$ is a proper conic, because then the two distinct points U and V on L can not be mapped onto the same point X. Therefore the conic $P(L)$ should be a real double line on S^2; but again, this is impossible. Note that we do not have to consider the case that $P(L)$ is a point on S^2, because in that case $P(r,s,t)$ would have two base points, and therefore be faithful. Hence we have shown that $P(r,s,t)$ is not a quadratic representation of the octant $\mathcal{T}$, whether faithful or not.

Finally, we show that $P(r,s,t)$ is not cubic. If $P(r,s,t)$ is of exactly cubic degree, then a general real line in the parameter domain, i.e. a line not passing through any base point of $P(r,s,t)$, is mapped onto a real cubic curve on S^2. Since a real cubic curve is always unbounded, but S^2 is bounded in E^3, this is a contradiction. □

Chapter 4

Interpolation on Quadric Surfaces by Rational Quadratic Spline Curves

In this chapter we discuss the problem of constructing a tangent continuous rational quadratic spline curve lying on a regular quadric surface $S : X^TAT = 0 \subset E^d$, $d \geq 3$, that interpolates a given point sequence $\{X_i\}_{i=1}^n$. It is shown that a necessary condition for the existence of a nontrivial interpolant is $(X_1^TAX_2)(X_i^TAX_{i+1}) > 0$, $i = 1, 2, \ldots, n-1$. Also considered is the Hermite interpolation problem on a quadric surface: a smooth interpolating curve on S is sought to interpolate two points and two associated tangent directions. The solution provided to this problem is a biarc consisting of two smoothly joining conic arcs on S, which is similar to the biarc interpolant in the plane or space [Bol75]. Several conditions for the existence of a biarc are derived, among which is a necessary and sufficient condition for the existence of a biarc whose two arcs are not major elliptic arcs. In addition, it is shown that this condition is always fulfilled on the sphere $S^{d-1} \subset E^d$ for generic interpolation data. An algorithm is presented to apply the biarc interpolant to solve the point interpolation problem by providing the tangent direction at each data point from nearby data points.

4.1 Introduction

4.1.1 Problems

Given a sequence of points $\{X_i\}_{i=1}^n$ on a quadric surface $S \subset E^d$, $d \geq 3$, we want to construct a smooth curve on S, i.e. a curve on S with unit tangent vector continuity, to interpolate $\{X_i\}_{i=1}^n$.

While it is not hard to identify this problem in CAGD, especially when $d = 3$, an interesting instance of this problem arises in computer animation, which involves the curve interpolation problem in the unit quaternion space that forms the unit sphere $S^3 \subset E^4$. More about this application can be found in [Sho85].

In this chapter we will mainly consider using the rational quadratic spline curve as the interpolating curve. Obviously, the rational quadratic spline is the simplest curve possible to solve the problem. There are many papers in the literature using rational quadratic splines, or conic splines, to solve interpolation and approximation problems in the plane or space, e.g. [Boo79, Pav83, Pra85]. Some authors studied special cases of rational quadratic splines, e.g. quadratic polynomial splines or circular arc splines. The advantages of rational quadratic splines are obvious. Its low algebraic degree ensures simple representation and evaluation, and in many applications its tangent continuity can well meet the requirement for smoothness. Of course, in general, the rational quadratic spline does not have curvature continuity in the plane [Far89] and torsion continuity in space, since conic sections are planar curves.

The first problem we will discuss is how to construct a smooth rational quadratic spline curve on S using a single segment between any two consecutive data points. This problem will be referred to as *point interpolation on the quadric surface*. We will show how to construct such a spline curve, and will prove that if the solution exists then all the line segments $\overline{X_i X_{i+1}}$, $i = 1, \ldots, n-1$, are on the same side of S and the spline has $d-2$ degrees of freedom. It will be seen that this spline curve is not locally controllable.

As the second problem, we will discuss the *biarc interpolant on the quadric surface*, or simply *biarc*, which is a composite curve consisting of two smoothly joining conic arcs on S and is made to interpolate two points X_0 and X_1 and two associated tangent directions on S. This problem arises in solving the interpolation problem in which, besides $\{X_i\}_{i=1}^n$, the tangent direction at each point X_i is also specified. Evidently, in general, the rational quadratic spline with a single curve segment between two consecutive points is insufficient to solve this problem due to the limited flexibility of conics. We will adopt an approach similar to the biarc interpolation in the plane [Bol75, Sab76], i.e. using a biarc on S to interpolate two consecutive points and the two end tangents, to match the degrees of freedom of the spline with the increased number of interpolation conditions. This approach provides an improved solution to the point interpolation problem on the quadric surface. Given a point sequence on a quadric surface, this improved solution is realized by generating the tangent direction at each point from nearby points.

Several conditions for the existence of these biarc interpolants will be derived. We mainly consider a special class of biarcs which do not contain major elliptic arcs, where a major elliptic arc is an arc greater than half an ellipse. For the existence of these kinds of biarcs a necessary and sufficient condition will be derived, and this condition will be shown to be satisfied for generic data on the sphere $S^{d-1} \subset E^d$, $d \geq 3$.

The remainder of this chapter is organized as follows. In the rest of this section we will review relevant preliminaries. Sections 4.2 and 4.3 deal with the two problems mentioned above. In Section 4.4 an algorithm is described to apply the biarc interpolant in the point interpolation problem. We conclude the discussion in Section 4.5 with a summary of these results and some open problems.

4.1.2 Preliminaries

Due to the nature of the problems discussed we will be mainly concerned with the Euclidean space E^d, $d \geq 3$, although frequently we have to use projective transformations. Therefore homogeneous coordinate representations for points in E^d will be used extensively. The homogeneous representation also simplifies the representation of different quadric surfaces to a unified form. The following remarks on calculation with homogeneous coordinates will be found useful.

A point in E^d is represented by one of its homogeneous coordinate representations $X = [x_1, \ldots, x_{d+1}]^T$, where $x_i \in \mathcal{R}$, the field of the real numbers, and at least one $x_i \neq 0$. Two $(d+1)$-tuples X_1, X_2 are defined to represent the same point in E^d if $X_1 = \rho X_2$ for some $\rho \neq 0$. If $x_{d+1} = 0$ then X is called a point at infinity with respect to E^d. The point represented by homogeneous coordinates X is denoted by $[X]$. When there is no danger of confusion we also call X a point.

We say that a finite point $X = [x_1, \ldots, x_{d+1}]^T \in E^d$ is in *normalized form* if $x_{d+1} = 1$. Let X,Y be two points in homogeneous form. Then the straight line XY is parameterized by $\lambda X + \mu Y$, where $\lambda, \mu \in \mathcal{R}$ are two independent real homogeneous parameters with $\lambda^2 + \mu^2 \neq 0$. When X,Y are in normalized homogeneous form, the straight line segment

$\overline{XY}$ is parameterized by $\lambda X + \mu Y$, where $\lambda, \mu \in \mathcal{R}$ are subject to $\lambda\mu \geq 0$ and $\lambda^2 + \mu^2 \neq 0$. A direction in E^d is represented the homogeneous coordinates T of a point at infinity. Note that T and $-T$ represent two opposite directions, though they stand for the same point at infinity. But T and ρT, where $\rho > 0$, represent the same direction.

Using homogeneous coordinates, a quadric surface $S \subset E^d$ is represented by $X^T AX = 0$, where A is a $(d+1) \times (d+1)$ real symmetric matrix, and X is a generic point on S in homogeneous representation. In this chapter we will mainly study real *regular* quadric surfaces. By *regular* we mean that the surface as a projective entity has no singular points. Regular quadrics are also said to be non-degenerate. The characterization for $X^T AX = 0$ to be a real regular quadric surface is that the matrix A is nonsingular and indefinite, i.e. $(X^T AX)(Y^T AY) < 0$ for some X and Y. A regular quadric surface is irreducible, i.e. not composed of hyperplanes [SeK52]. For a regular quadric surface S, the tangent hyperplane of S at a point X_0, $X_0^T AX = 0$, which is the polar plane of X_0 with respect to S, is well defined, because when A is nonsingular $X_0^T A \neq 0$ if $X_0 \neq 0$. Like in E^3, a straight line is called a *generating line* of S if it is entirely contained in S. It is easily verified that two distinct points X_0,X_1 on S are on the same generating line of S, i.e. $(\lambda X_0+\mu X_1)^T A(\lambda X_0+\mu X_1) = 0$ for all $(\lambda, \mu) \neq (0,0)$, if and only if $X_0^T AX_1 = 0$. Two points X_0,X_1 are defined to be on *opposite sides* of $S : X^T AX = 0$, if $(X_0^T AX_0)(X_1^T AX_1) < 0$; they are on *the same side* of S if $(X_0^T AX_0)(X_1^T AX_1) > 0$. Of course, $X_0^T AX_0 = 0$ means $X_0 \in S$.

A quadric surface $S \subset E^d$ is composed of at most two disjoint real connected components; for if there are more than two, then three distinct points can be chosen on S, each from a different component. If these three points are collinear, the line passing through them intersects S in three points. Then by Bezout's Theorem, the line lies entirely in S, connecting the three components; this is a contradiction. If the three points are not collinear, then the conic section obtained as the intersection of S and the plane determined by the three points has at least three real components, which is again impossible. Quadric surfaces with one and two components are exemplified by the ellipsoid and hyperboloid of two sheets in E^3, respectively. A necessary requirement for two points on S to be interpolated by a continuous curve on S is that they should be on the same component of S. Throughout we will assume that the points $\{X_i\}_{i=1}^n$ to be interpolated are on the same component of S.

When two surfaces can be mapped into each other by a projective transformation we say that they are projectively equivalent, otherwise projectively different. Similarly we have the concept of affine equivalence. There are five classes of projectively different real quadric surfaces in E^3: they are, taking one representative from each class, the elliptic paraboloid, the hyperbolic paraboloid, the quadric cone, the parallel planes and the coincident planes.

Only the first two kinds of surfaces are regular. The last two kinds of surfaces are of no interest to us. Each of the first three classes can be further divided into subclasses of affinely equivalent surfaces. The first class contains the ellipsoid, the hyperboloid of two sheets, and the elliptic paraboloid; the second class contains the hyperboloid of one sheet and the hyperbolic paraboloid; the third class, often referred to as *properly degenerate quadric surfaces*, contains the quadric cone, the elliptic cylinder, the parabolic cylinder, and the hyperbolic cylinder.

The properly degenerate quadric surfaces are not regular surfaces; for cylinders the singular point is at infinity. Most of our results in E^d will be derived only for regular quadric surfaces, but we will be explicit when the results can be extended to any singular quadric surface.

The rational quadratic Bézier curve will be used to represent each piece of a spline in our discussion. Here, for the purpose of this chapter and the next, we will discuss the homogeneous representation of the rational quadratic Bézier curve

$$P(u) = w_0 P_0 B_{0,2}(u) + w_1 P_1 B_{1,2}(u) + w_2 P_2 B_{2,2}(u), \quad 0 \le u \le 1, \tag{4.1}$$

where P_0, P_1 and P_2, which are in homogeneous representation, are called the *control points* of the curve; the w_i's are the *weights* associated with the P_i's, $i = 0, 1, 2$; $B_{i,2}(t) = \frac{2!}{i!(2-i)!}(1-u)^{2-i}u^i$, $i = 0, 1, 2$, are the *second degree Bernstein polynomials.* Here note that P_0, P_1 and P_2 are not the control points in the usual sense since different homogeneous representations P_i of the $[P_i]$ may result in different curves.

The following notation is about conics and a finite piece of a conic. A conic that is composed of straight lines is said to be *degenerate*, otherwise *proper.* A *conic arc* refers to a continuous and finite piece of a conic section. A *smooth conic arc* refers to a tangent continuous conic arc, with a line segment included even though the underlying conic is degenerate. A *proper conic arc* refers to a conic arc on a proper conic; therefore a proper conic arc is a smooth conic arc. A proper conic arc determines a plane, since the underlying conic does. A continuous and finite piece of a rational quadratic curve is called a *conic arc*, or interchangeably, a *rational quadratic segment.* Through a rational linear reparameterization, a conic arc can be represented in the standard form [Pat86]

$$P(u) = P_0 B_{0,2}(u) + w P_1 B_{1,2}(u) + P_2 B_{2,2}(u), \quad u \in [0, 1], \tag{4.2}$$

where P_0, P_2 are in normalized homogeneous form; w is called the *weight* of the curve, which is not necessarily positive. Note that w is not unique since different representations of $[P_1]$ come into play. If the representation P_1 is fixed, two weights with opposite signs give complementary arcs of the same conic section [Lee87]. Therefore both arcs are continuous if and only if the conic is an ellipse.

A curve segment (4.2) is continuous provided $wx_{d+1} \geq 0$, where x_{d+1} is the last component of P_1. When $w = 0$ the curve becomes the line segment $\overline{P_0P_2}$; when $x_{d+1} = 0$ and $w \neq 0$ the curve is half an ellipse. In the following we will mainly consider the case $w \neq 0$, as it will be shown later on that straight line segments are essentially useless in our solution.

Definition 4.1.1: *Suppose that P_1 in (4.2) is in normalized form when it is a finite point. A real, nonzero, and finite weight w is called a* positive weight *if $w > 0$; it is called a* negative weight *otherwise.*

Note that a real nonzero finite weight w is either a positive weight or a negative weight; when $[P_1]$ is a point at infinity, the positive weight and the negative weight are defined with respect to the particular representation P_1. A rational quadratic curve with a positive weight can only give a continuous piece of a hyperbolic arc, parabolic arc or minor elliptic arc, including a semi-ellipse when P_1 is at infinity. An important property of the rational quadratic Bézier curve is that the straight line segments $\overline{P_0P_1}$, $\overline{P_1P_2}$ are tangent to the curve $P(u)$ at $P_0 = P(0)$, $P_2 = P(1)$, respectively.

The condition on the smooth joining of two rational quadratic Bézier curves is important in our discussion. Let the control polygon of two Bézier segments be $X_0Y_0X_1$ and $X_1Y_1X_2$ respectively, where Y_0, Y_1 may be points at infinity. It is evident that for the two Bézier curves to join smoothly the three points Y_0, X_1 and Y_1 have to be collinear with $[X_1] \neq [Y_0]$, $[X_1] \neq [Y_1]$. In this case the following two rules are easily observed, which are given as a lemma for later reference.

Lemma 4.1.1: *When the joint point X_1 is between Y_0 and Y_1 the two Bézier curves join smoothly if and only if they both simultaneously have positive a weight or a negative weight. When X_1 is not between Y_0 and Y_1, including the case $[Y_0] = [Y_1]$, the two Bézier curves join smoothly if and only if one curve has a positive weight and the other has a negative weight.*

For the above rules to make sense it is required that the arc given by a negative weight is continuous.

4.2 Point interpolation on a quadric

4.2.1 Local representation

Let $\{X_i\}_{i=1}^n$, $n \geq 3$, be a point sequence in normalized form on a quadric surface S: $X^T AX = 0 \in E^d$, $d \geq 3$. Assume that $\{X_i\}_{i=1}^n$ are on the same real component of S in the case that S is composed of two components in E^d. Our goal in this section is to construct a smooth rational quadratic spline on S to interpolate $\{X_i\}_{i=1}^n$. First we consider the existence and properties of a single quadratic Bézier curve on S interpolating two points, say X_0 and X_1.

Let the tangent hyperplanes of S at X_0 and X_1 be Q_0: $X_0^T AX = 0$ and Q_1: $X_1^T AX = 0$, respectively. Let L_0 be the intersection of Q_0, Q_1. Then L_0 is a $(d-2)$-dimensional affine manifold; L_0 is a straight line when $d = 3$. Let $C_0 : P_0(u)$ be a rational quadratic Bézier curve on S interpolating X_0 and X_1, and let X_0, Y_0 and X_1 be the control points of $P_0(u)$. Then, as C_0 is on S, it is necessary for Y_0 to be on L_0, i.e. $X_0^T AY_0 = 0$ and $X_1^T AY_0 = 0$; for otherwise the straight line Y_0X_0 or Y_0X_1 would not be tangent to S.

Suppose that $C_0 : P_0(u)$ is in the standard Bézier representation

$$P_0(u) = X_0B_{0,2}(u) + wY_0B_{1,2}(u) + X_1B_{2,2}(u), \quad 0 \leq u \leq 1. \tag{4.3}$$

The weight w must satisfy $P_0(u)^T AP_0(u) = 0$ for all $u \in [0,1]$ since $C_0 \subset S$. Substituting (4.3) in $P_0(u)^T AP_0(u) = 0$, noting that $X_0^T AX_0 = X_1^T AX_1 = X_0^T AY_0 = X_1^T AY_0 = 0$, and $B_{1,2}^2(u) = 4B_{0,2}(u)B_{2,2}(u)$, we obtain

$$2X_0^T AX_1B_{0,2}(u)B_{2,2}(u) + w^2Y_0^T AY_0B_{1,2}^2(u) = 0,$$

or when $Y_0^T AY_0 \neq 0$

$$w^2 = -\frac{X_0^T AX_1}{2Y_0^T AY_0}. \tag{4.4}$$

When the right hand side of (4.4) is finite and nonnegative, (4.4) is said to be *valid.* A real value of the weight can be calculated from (4.4) if and only if (4.4) is valid. Note that (4.4) being valid is just a necessary condition for $P_0(u)$ to be a conic arc on S interpolating X_0 and X_1; a necessary and sufficient condition is that $Y_0 \in L_0$ and that (4.4) is valid. Now we shall find a condition on Y_0 under which (4.4) is valid since only real values of w yield real curves. First we exclude a trivial case.

Lemma 4.2.1: *Let X_0, X_1 be distinct points on the same generating line of the quadric surface S. Then the line segment $\overline{X_0X_1}$ is the only smooth conic arc on S interpolating X_0,X_1.*

PROOF: We just need to show that there is no other proper conic arc on S interpolating X_0 and X_1. Suppose there is such a proper conic C_0 passing through X_0,X_1. Then the two dimensional plane containing C_0 intersects the surface S in C_0 plus the straight line X_0X_1, contradicting the fact that any plane section of a quadric surface is a conic. □

Obviously, X_0, X_1 are on the same generating line of S if and only if $[X_0 + X_1] \in S$, which is a point on the segment $\overline{X_0X_1}$ since X_0 and X_1 are in normalized form. But $(X_0 + X_1)^T A(X_0 + X_1) = 0$ is equivalent to $X_0^T AX_1 = 0$. Thus we have

Lemma 4.2.2: *Let w be determined by (4.4) and $Y_0^T AY_0 \neq 0$. Then $w = 0$ if and only if X_0,X_1 are on the same generating line of S.*

Since $w = 0$ in (4.3) gives a straight line segment $P_0(u)$, Lemma 4.2.2 is just a restatement of Lemma 4.2.1. Because of the following fact, the case in which two consecutive points X_0,X_1 are on the same generating line of S is not of interest to us, and will therefore be excluded.

Lemma 4.2.3: *On a regular quadric surface S a straight line segment and a proper conic arc can not meet with tangent continuity.*

PROOF: The proof is very much like that of Lemma 4.2.1. Suppose that a proper conic arc C and a straight line segment l on S join with common tangent T. Let P_C be the two dimensional plane determined by C. Then P_C contains T, and therefore contains l. So the plane P_C intersects the quadric surface S in a proper conic that contains C plus the straight line containing l. This absurdity indicates that the lemma must hold. □

The following theorem provides a geometric condition for the existence of and a way of constructing the local interpolating rational quadratic curve.

Theorem 4.2.4: *Let X_0,$X_1 \in S$: $X^T AX = 0 \subset E^d$ be distinct points on the same component but not on the same generating line of S. Then X_0,Y_0 and X_1 form the control points of a smooth rational quadratic Bézier curve on S interpolating X_0,X_1 if and only if $X_0^T AY_0 = X_1^T AY_0 = 0$ and $(Y_0^T AY_0)(X_0^T AX_1) < 0$, or geometrically, $Y_0 \in L_0$ and Y_0 and the line segment $\overline{X_0X_1}$ are on opposite sides of S.*

PROOF: Let X_0 and X_1 be in normalized form. Then $X_0 + X_1$ stands for a point on the line segment $\overline{X_0X_1}$. First we remark that the segment $\overline{X_0X_1}$ is entirely on the same side of S as the point $[X_0 + X_1]$, since X_0 and X_1 are the only intersections of the straight line X_0X_1 with S.

When X_0,Y_0 and X_1 are the control points of a smooth Bézier curve of the form (4.3) on S, obviously $Y_0 \in L_0$, i.e. $X_0^T A Y_0 = X_1^T A Y_0 = 0$. Also X_0, Y_0 and X_1 are not collinear; for otherwise, the Bézier curve would be a generating line of S passing through X_0 and X_1. Also we have $Y_0^T A Y_0 \neq 0$. For otherwise, from $Y_0^T A Y_0 = 0$, i.e. $Y_0 \in S$, combined with $X_0^T A Y_0 = X_1^T A Y_0 = 0$, it would follow that X_0 and Y_0 are on the same generating line of S, and so are X_1 and Y_0 for the same reason. Since the intersection of the quadric and the plane containing the rational Bèzier curve already contains two lines, these must be the conic of intersection. Hence the Bézier curve would consist of the joint line segments $\overline{X_0Y_0}$ and $\overline{Y_0X_1}$, which is not smooth. Thus $Y_0^T A Y_0 \neq 0$. Since the curve is well defined, $-X_0^T A X_1 / (2 Y_0^T A Y_0) = w^2 \geq 0$. As X_0,X_1 are not on the same generating line of S, $X_0^T A X_1 \neq 0$; therefore $(X_0^T A X_1)(Y_0^T A Y_0) < 0$. Since $(X_0 + X_1)^T A (X_0 + X_1) = 2 X_0^T A X_1$, we have

$$[(X_0 + X_1)^T A (X_0 + X_1)](Y_0^T A Y_0) < 0.$$

Hence Y_0 and the segment $\overline{X_0X_1}$ are on the opposite sides of S.

Now suppose that (i) $X_0^T A Y_0 = X_1^T A Y_0 = 0$ and (ii) $(Y_0^T A Y_0)(X_0^T A X_1) < 0$. Since $Y_0 \in L_0$, by (i), we can construct a Bézier curve on S of the form (4.3), with the weight w determined by (4.4). By (ii), we have

$$\frac{-X_0^T A X_1}{2 Y_0^T A Y_0} > 0.$$

Therefore a positive weight w can be solved for from (4.4), i.e. X_0,Y_0 and X_1 are the control points of a smooth Bézier curve on S interpolating X_0 and X_1. □

It is evident that when X_0,X_1 are distinct points on the sphere $S^{d-1} \subset E^d$ or any surface that is affinely equivalent to S^{d-1}, $(Y_0^T A Y_0)(X_0^T A X_1) < 0$ holds for any $Y_0 \in L_0$. Therefore we have

Lemma 4.2.5: *Let X_0, $X_1 \in S^{d-1} \subset E^d$ be distinct points. Then for any $Y_0 \in L_0$, X_0, Y_0 and X_1 are the control points of two smooth rational quadratic Bézier curves interpolating X_0 and X_1 on S^{d-1}, one with the positive weight and the other with the negative weight.*

In fact, Lemma 4.2.5 can be generalized to any properly degenerate quadric surface in E^3, i.e. any proper cylinder or the quadric cone, with one slight modification. When S is a properly degenerate quadric surface in E^3 and X_0, X_1 are not on the same generating line of S, $(Y_0^T A Y_0)(X_0^T A X_1) < 0$ for any $Y_0 \in L_0$, with the one exception that Y_0 is not the vertex of a quadric cone. Here for convenience we say that the quadric cone in E^3 is composed of two components joining at the vertex. This statement is not strictly accurate by definition but is based rather on the observation that if two points are on different parts of a quadric cone S then either they are on the same generating line of S or the condition $(Y_0^T A Y_0)(X_0^T A X_1) < 0$ in Theorem 4.2.4 is violated for any $Y_0 \in L_0$.

For a general regular quadric surface we have only a weaker result. From Theorem 4.2.4 it is seen that $Y_0 \in L_0$ gives a smooth interpolating Bézier curve (4.3) if and only if (4.4) is valid.

Lemma 4.2.6: *Let X_0, X_1 be distinct points on the same component but not on the same generating line of a regular quadric surface S: $X^TAX = 0 \in E^d$, $d \geq 3$. Then there exists $Y_0 \in L_0$ such that (4.4) is valid.*

Lemma 4.2.6 is equivalent to the statement that for any two distinct points on the same component of a regular quadric S in E^d, there exists a smooth conic arc on S connecting the two points. To prove Lemma 4.2.6, we need the following lemma.

Lemma 4.2.7: *Let A be a real $n \times n$ nonsingular symmetric matrix. Let A have p positive and r negative eigenvalues, $p + r = n$. Let B be an $n \times (n-1)$ matrix of rank $n-1$. Then the symmetric matrix B^TAB has at least $p-1$ positive and at least $r-1$ negative eigenvalues.*

PROOF: Since the rank of B is $n-1$, we can add a new column b to it such that $D = [B, b]$ is nonsingular. Then B^TAB is the leading $(n-1) \times (n-1)$ principal submatrix of D^TAD. By the Sylvester law of inertia [GoV89, pp. 416–417], the number of positive eigenvalues and the number of negative eigenvalues of D^TAD are the same as those of A. Let the eigenvalues of B^TAB be λ_i, $i = 1, 2, \ldots, n-1$, and let the eigenvalues of D^TAD be μ_j, $j = 1, 2, \ldots, n$. Since the eigenvalues of B^TAB separate those of D^TAD [Wil65, pp. 103–104], i.e.

$$\mu_1 \leq \lambda_1 \leq \mu_2 \leq \lambda_2 \leq \cdots \leq \lambda_{n-1} \leq \mu_n,$$

we conclude that B^TAB has at least $p-1$ positive eigenvalues and at least $r-1$ negative eigenvalues.□

PROOF of Lemma 4.2.6: First we need an affine classification of real quadric surfaces in E^d [Xu65, pp. 471–474]. It is straightforward to show that any real regular quadric surface in E^d is affinely equivalent to one of the following forms:

1. $X^TAX = 0$, where $A = \text{diag}[1, \sigma_2, \ldots, \sigma_d, -1]$, $\sigma_i = \pm 1$, $i = 2, \ldots, d$; or
2. $X^TAX = 0$, where $A = \text{diag}\left[\sigma_1, \ldots, \sigma_{d-1}, \begin{bmatrix} 0 & -1 \\ -1 & 0 \end{bmatrix}\right]$, $\sigma_i = \pm 1$, $i = 1, \ldots, d-1$.

Let A have p positive and r negative eigenvalues. Since A is indefinite for $X^TAX = 0$ to be a real surface, $p \geq 1$ and $r \geq 1$.

We just have to show that the lemma holds for surfaces in these two canonical forms. First consider the class of quadrics $X^TAX = 0$ with $p = 1$ or $r = 1$. These quadrics must be in one of the following three cases:

1. $X^TAX = 0$ with $A = \mathrm{diag}[I_d, -1]$;
2. $X^TAX = 0$ with $A = \mathrm{diag}[-1, I_d]$, which gives the same quadric as by $A = \mathrm{diag}[1, -I_d]$;
3. $X^TAX = 0$ with $A = \mathrm{diag}\left[I_{d-1}, \begin{bmatrix} 0 & -1 \\ -1 & 0 \end{bmatrix}\right]$.

For these three cases, $p = d$ and $r = 1$. Define a point X_0 to be inside $S : X^TAX = 0$ if $X_0^TAX_0 < 0$. Let S be any one of the above three surfaces. Then, given any two distinct real points $X_0 = [x_{0,1}, \ldots, x_{0,d}, 1]^T$, $X_1 = [x_{1,1}, \ldots, x_{1,d}, 1]^T$ on the same component of S, it will be shown that the line segment $\overline{X_0X_1}$ is inside S.

(1) The case $A = \mathrm{diag}[I_d, -1]$:

$$(X_0 + X_1)^TA(X_0 + X_1) = 2X_0^TAX_1 = -(X_0 - X_1)^TA(X_0 - X_1) < 0.$$

(2) The case $A = \mathrm{diag}[-1, I_d]$: Since $x_1 = 0$ is the separating hyperplane of S, S has two components. Since X_0, X_1 are on the same component of S, we have $x_{0,1}x_{1,1} > 0$. Then

$$\begin{aligned} & x_{0,1}x_{1,1}(X_0 + X_1)^TA(X_0 + X_1) = 2x_{0,1}x_{1,1}X_0^TAX_1 \\ = \; & -(x_{1,1}X_0 - x_{0,1}X_1)^TA(x_{1,1}X_0 - x_{0,1}X_1) < 0. \end{aligned}$$

Thus $(X_0 + X_1)^TA(X_0 + X_1) < 0$.

(3) The case $A = \mathrm{diag}\left[I_{d-1}, \begin{bmatrix} 0 & -1 \\ -1 & 0 \end{bmatrix}\right]$: Since X_0, X_1 are real points on S, $x_{0,d} \geq 0$ and $x_{1,d} \geq 0$. Then

$$\begin{aligned} & -(x_{0,d}+1)(x_{1,d}+1)(X_0 + X_1)^TA(X_0 + X_1) = -2(x_{0,d}+1)(x_{1,d}+1)X_0^TAX_1 \\ = \; & [(x_{1,d}+1)X_0 - (x_{0,d}+1)X_1]^TA[(x_{1,d}+1)X_0 - (x_{0,d}+1)X_1] \\ = \; & \sum_{i=1}^{d-1}[(x_{1,d}+1)x_{0,i} - (x_{0,d}+1)x_{1,i}]^2 - 2[(x_{1,d}+1)x_{0,d} - (x_{0,d}+1)x_{1,d}][(x_{1,d}+1) - (x_{0,d}+1)] \\ = \; & \sum_{i=1}^{d-1}[(x_{1,d}+1)x_{0,i} - (x_{0,d}+1)x_{1,i}]^2 + 2(x_{1,d} - x_{0,d})^2 > 0, \end{aligned}$$

since X_0, X_1 are distinct points. Thus $(X_0 + X_1)^TA(X_0 + X_1) < 0$ since $x_{0,d} + 1 > 0$ and $x_{1,d} + 1 > 0$. Hence for the three quadrics the segment $\overline{X_0X_1}$ is inside the surface.

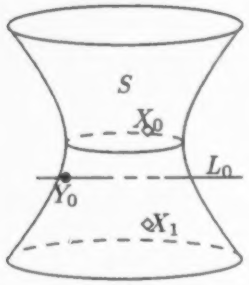


Fig. 4.2.1 Local interpolation on hyperboloid of one sheet: The straight line L_0 intersects the hyperboloid of one sheet in two points. Therefore a point $Y_0 \in L_0$ can be selected that is on the same side of the surface as the segment $\overline{X_0X_1}$.

Let $L = Q_0 \bigcap Q_1$, where Q_0 and Q_1 are the tangent hyperplanes of S at X_0 and X_1, given by $X_0^T AX = 0$ and $X_1^T AX = 0$, respectively. Let Y_i, $i = 1, \ldots, d-1$, be $d-1$ affinely independent points in L. Similar to the Gram-Schmidt orthogonalization process [GoV89, p. 218], the Y_i can be constructed so that $Y_i^T AY_j = 0$ for $i \neq j$. Let $Z = \lambda X_0 + \mu X_1$ be a variable point on the straight line X_0X_1. Then $Z^T AY_i = 0$ for $i = 1, \ldots, d-1$. Let $B = [Y_1, \ldots, Y_{d-1}, Z]$. Then $B^T AB = \mathrm{diag}[Y_1^T AY_1, \ldots, Y_{d-1}^T AY_{d-1}, Z^T AZ]$. So B has rank d. By Lemma 4.2.7, $B^T AB$ has at least $d-1$ positive eigenvalues since A has d positive eigenvalues. But since Z is on the straight line X_0X_1, it can be chosen so that $Z^T AZ < 0$; therefore the $Y_i^T AY_i > 0$. Thus $Y^T AY > 0$ for any $Y \in L$. Hence when S is any of the three quadrics, for any $Y \in L$, the line segment $\overline{X_0X_1}$ and Y are on opposite sides of S.

Now consider the remaining case, i.e. the quadrics $X^T AX = 0$ with $p \geq 2$ and $r \geq 2$. Let $B = [Y_1, \ldots, Y_{d-1}, Z]$ be the same matrix as constructed above. For the same reason, B has rank d and $B^T AB = \mathrm{diag}[Y_1^T AY_1, \ldots, Y_{d-1}^T AY_{d-1}, Z^T AZ]$. In this case, by Lemma 4.2.7, $B^T AB$ has at least one positive eigenvalue and one negative eigenvalue, so it is indefinite. Therefore the $Y_i^T AY_i$ do not have the same sign; for otherwise, choosing $Z^T AZ$ to have the same sign as the $Y_i^T AY_i$, $B^T AB$ would become positive or negative definite, a contradiction. Since the $Y_i^T AY_i$ have different signs, there exists $Y \in L$ such that $\overline{X_0X_1}$ and Y are on opposite sides of S. □

Lemma 4.2.6 can not be made as strong as Lemma 4.2.5 because on a hyperboloid of one sheet S in E^3 it is easy to give two points X_0, $X_1 \in S$ and a $Y_0 \in L_0$ such that Y_0 and $\overline{X_0X_1}$ are on the same side of S. See Fig. 4.2.1. Thus by Theorem 4.2.4, X_0, Y_0 and X_1 can not be the control points of a well defined interpolating curve (4.3).

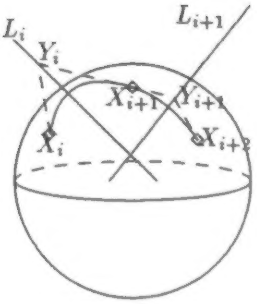


Fig. 4.2.2 Control points: The control point $Y_{i+1} \in L_{i+1}$ is the projection of $Y_i \in L_i$ through X_{i+1}.

4.2.2 Construction of interpolating splines

Given $\{X_i\}_{i=1}^n$, $n \geq 3$, on S: $X^T A X = 0 \subset E^d$, now we consider constructing a smooth rational quadratic spline on S interpolating $\{X_i\}_{i=1}^n$. Let X_i, Y_i and X_{i+1} be the control points of the local curve segment C_i in the standard Bézier representation (4.3). We need to determine the Y_i's such that C_i and C_{i+1} join with tangent continuity, $i = 1, \ldots, n-1$. We assume that any two consecutive points are distinct and $\{X_i\}_{i=1}^n$ are on the same component of S. By Lemma 4.2.1, when $\{X_i\}_{i=1}^n$ are all on the same generating line of S, either there is no solution or that generating line is a trivial solution. Therefore we will assume that not all $\{X_i\}_{i=1}^n$ are on the same generating line of S. And by Lemma 4.2.3, we may further assume that no two consecutive points are on the same generating line of S.

Now given $\{X_i\}_{i=1}^n$, by Lemma 4.2.6, we can first choose $Y_1 \in L_1$ such that (4.4) is valid. Suppose, recursively, that Y_i has been determined. Let us find Y_{i+1}. Since $\overline{Y_i X_{i+1}}$ and $\overline{Y_{i+1} X_{i+1}}$ are the tangents to C_i and C_{i+1} at their joint point X_{i+1}, in order for C_i and C_{i+1} to have common tangent at X_{i+1}, the point $Y_{i+1} \in L_{i+1}$ has to be the projection of $Y_i \in L_i$ through X_{i+1}. See Fig. 4.2.2 for an illustration in the case $d = 3$. So, inductively, Y_{i+1} depends projectively on Y_1. The following lemma gives an expression for this dependence.

Lemma 4.2.8: *Let* $M_i = \prod_{j=1}^i R_j$, *with* $R_1 = I$, *the identity matrix, and*

$$R_j = X_j X_{j+1}^T A - (X_j^T A X_{j+1}) I, \quad j = 2, \ldots, n-1.$$

If an interpolating quadratic spline exists, then $Y_i = M_i Y_1$, $i = 1, \ldots, n-1$.

PROOF: Because Y_i, X_{i+1} and Y_{i+1} are collinear, we have

$$Y_{i+1} = aX_{i+1} + bY_i,$$

for some constants a and b. Premultiplying $X_{i+2}^T A$ to both sides, we obtain

$$0 = a(X_{i+2}^T AX_{i+1}) + b(X_{i+2}^T AY_i),$$

since $X_{i+2}^T AY_{i+1} = 0$. So omitting a nonzero multiplicative constant, we can put

$$\begin{aligned} Y_{i+1} &= (X_{i+2}^T AY_i)X_{i+1} - (X_{i+2}^T AX_{i+1})Y_i \\ &= [X_{i+1}X_{i+2}^T A - (X_{i+1}^T AX_{i+2})I]Y_i. \end{aligned} \tag{4.5}$$

Let $R_j = X_j X_{j+1}^T A - (X_j^T AX_{j+1})I$, $j = 2, \ldots, n-1$. Then the theorem follows. □

The following theorem is the main result of this section which gives a necessary condition for the existence of an interpolating spline to the points $\{X_i\}_{i=1}^n$.

Theorem 4.2.9: *Let a sequence of points $\{X_i\}_{i=1}^n$ be given on the same component of a quadric surface S: $X^T AX = 0 \in E^d$. Assume that no two consecutive points X_i, X_{i+1} are on the same generating line of S. A necessary condition for the existence of a smooth rational quadratic spline curve on S interpolating $\{X_i\}_{i=1}^n$ is that $(X_1^T AX_2)(X_i^T AX_{i+1}) > 0$, $i = 1, \ldots, n-1$, i.e. all segments $\overline{X_i X_{i+1}}$ are on the same side of S.*

The next lemma will be used in the proof of the above theorem and later on.

Lemma 4.2.10: *Let $\{X_i\}_{i=1}^n$ be given as in Theorem 4.2.9. Let $Y_1 \in L_1$ and $Y_i = M_i Y_1$, $i = 2, \ldots, n-1$, as defined in Lemma 4.2.8. If $Y_1^T AY_1 \neq 0$, then $(Y_1^T AY_1)(Y_i^T AY_i) > 0$, $i = 1, 2, \ldots, n-1$, i.e. all Y_i's are on the same side of S.*

PROOF: As obtained in the proof of Lemma 4.2.8,

$$Y_{i+1} = (X_{i+2}^T AY_i)X_{i+1} - (X_{i+1}^T AX_{i+2})Y_i, \quad i = 1, \ldots, n-2.$$

Since $X_{i+1}^T AX_{i+1} = X_{i+1}^T AY_i = 0$, it follows from the above expression that

$$Y_{i+1}^T AY_{i+1} = (X_{i+1}^T AX_{i+2})^2 (Y_i^T AY_i).$$

Since X_{i+1}, X_{i+2} are not on the same generating line of S, $(X_{i+1}^T AX_{i+2})^2 > 0$. Hence the lemma is proved inductively. □

PROOF of Theorem 4.2.9: The existence of the smooth interpolating spline implies that (4.4) is valid for all $Y_i = M_i Y_1 \in L_i$, $i = 1, \ldots, n-1$, i.e. $Y_i^T AY_i \neq 0$ and $(X_i^T AX_{i+1})/(Y_i^T AY_i) < 0$, since $X_i^T AX_{i+1} \neq 0$. By Lemma 4.2.10, all the $Y_i^T AY_i$ have the same sign. Therefore all the $X_i^T A X_{i+1}$ have the same sign, i.e. $(X_1^T A X_2)(X_i^T A X_{i+1}) > 0$. Since $(X_i + X_{i+1})^T A (X_i + X_{i+1}) = 2X_i^T A X_{i+1}$, all segments $\overline{X_i X_{i+1}}$ are on the same side of S.□

The condition given in Theorem 4.2.9 is, in general, not sufficient, although it can be shown that (4.4) is valid for each i when the condition is fulfilled. When the alignment of the data $\{X_i\}_{i=1}^n$ on S has radical changes, the weights found from (4.4) must be determined by applying Lemma 4.1.1 to enforce tangent continuity between adjacent curve segments. Since some weights may not be positive, there may result in a discontinuous Bézier curve, as we have remarked that the arc corresponding to a negative weight is continuous only when the underlying conic is elliptic. This analysis indicates the following result.

Theorem 4.2.11: *Given a point sequence $\{X_i\}_{i=1}^n$ on $S \subset E^d$ which is affinely equivalent to the sphere S^{d-1}, and any point $Y_1 \in L_1$, there exists a smooth rational quadratic spline on S interpolating $\{X_i\}_{i=1}^n$, with the initial control point Y_1 as given.*

PROOF: Since the condition in Theorem 4.2.9 is satisfied when $\{X_i\}_{i=1}^n \subset S$, we just need to show that in this case it is also sufficient.

Let $Y_1 \in L_1$ and Y_i, $i = 2, \ldots, n-1$ be given as in Lemma 4.2.8. By Lemma 4.2.5, for any $Y_1 \in L_1$ we have $-(X_1^T A X_2)/(2Y_1^T A Y_1) > 0$ since $X_1^T A X_2 \neq 0$. By Lemma 4.2.10, $(Y_i^T A Y_i)(Y_1^T A Y_1) > 0$, $i = 1, 2, \ldots, n-1$. Since $\{X_i\} \subset S$ and $X_i \neq X_{i+1}$, $(X_1^T A X_2)(X_i^T A X_{i+1}) > 0$. Therefore $-(X_i^T A X_{i+1})/(2Y_i^T A Y_i) > 0$, $i = 1, 2, \ldots, n-1$, i.e. two real weights can be obtained from (4.4) for each i, any one of which yields a continuous and smooth Bézier curve segment since the underlying curve is an ellipse. So the required interpolating spline is given by applying Lemma 4.1.1 to choose the appropriate weights successively to ensure tangent continuity between all adjacent C_i. □

We remark that the condition of Theorem 4.2.9 imposes a substantial restriction on other kinds of quadric surfaces. For example, on a hyperboloid of one sheet in E^3, it is easy to come up with a point sequence $\{X_i\}_{i=1}^n$ such that not all the segments are on the same sides of the surface. See Fig. 4.2.3. Hence by Theorem 4.2.9 it is impossible to construct a smooth rational quadratic spline curve on the surface to interpolate $\{X_i\}_{i=1}^n$.

Fig. 4.2.4 illustrates the application of the method discussed above in interpolating six data points on a sphere in E^3 by choosing different points $Y_1 \in L_1$. We still do not know what the best choice of Y_1 is or if there is always an acceptable choice of Y_1 for all possible data. In general, the position of Y_1 has an undesirable global influence over the whole curve. Later in this thesis we will use biarc interpolants to solve this problem.

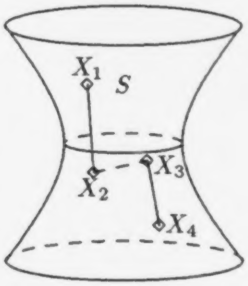


Figure 4.2.3 Data on hyperboloid of one sheet: All the line segments between consecutive points of the given sequence are not on the same side of the hyperboloid of one sheet S. Therefore this sequence does not admit a rational quadratic spline interpolant on S.

4.2.3 Closed interpolating splines

Given points $\{X_i\}_{i=1}^{n+2}$, $n \geq 3$ on S: $X^TAX = 0$ with $X_{n+1} = X_1$, $X_{n+2} = X_2$, we now consider constructing a closed smooth rational quadratic spline on S interpolating $\{X_i\}_{i=1}^{n}$. In order for this problem to have a solution, clearly it is necessary that there exist $Y_1 \in L_1$ such that $M_{n+1}Y_1 = \rho Y_1$ for some $\rho \neq 0$, where M_{n+1} is defined in Lemma 4.2.8.

From its definition in Lemma 4.2.8, $M_i = \prod_{j=1}^{i} R_j$, where R_j induces a projection from L_{j-1} to L_j. Therefore M_i, when restricted to L_1, is a projective transformation from L_1 to L_i; in particular, M_{n+1} induces a projective transformation on L_1. Thus the following theorem is evident.

Theorem 4.2.12: *The closed interpolation problem has a solution if and only if there exists a smooth rational quadratic spline interpolating $\{X_i\}_{i=1}^{n+2}$ with initial control point $Y_1 \in L_1$ such that Y_1 is a real fixed point of M_{n+1}.*

Now in arbitrary dimension precise conditions for the existence of a real fixed point of M_{n+1} in L_1 are still not known. We will discuss the two most encountered cases, that is $d = 3$ and $d = 4$. When $d = 3$, L_1 is a straight line in E^3, and M_{n+1} induces a homography $H(L_1)$ on L_1. *A homography* on a straight line is a rational linear transformation on it. *A united point* of a homography is one of its fixed points on the straight line. By the theory of homographies on straight lines [SeK52, p. 50], $H(L_1)$ has either two distinct real united points, or a double real united point, or a pair of conjugate complex united points. So M_{n+1}

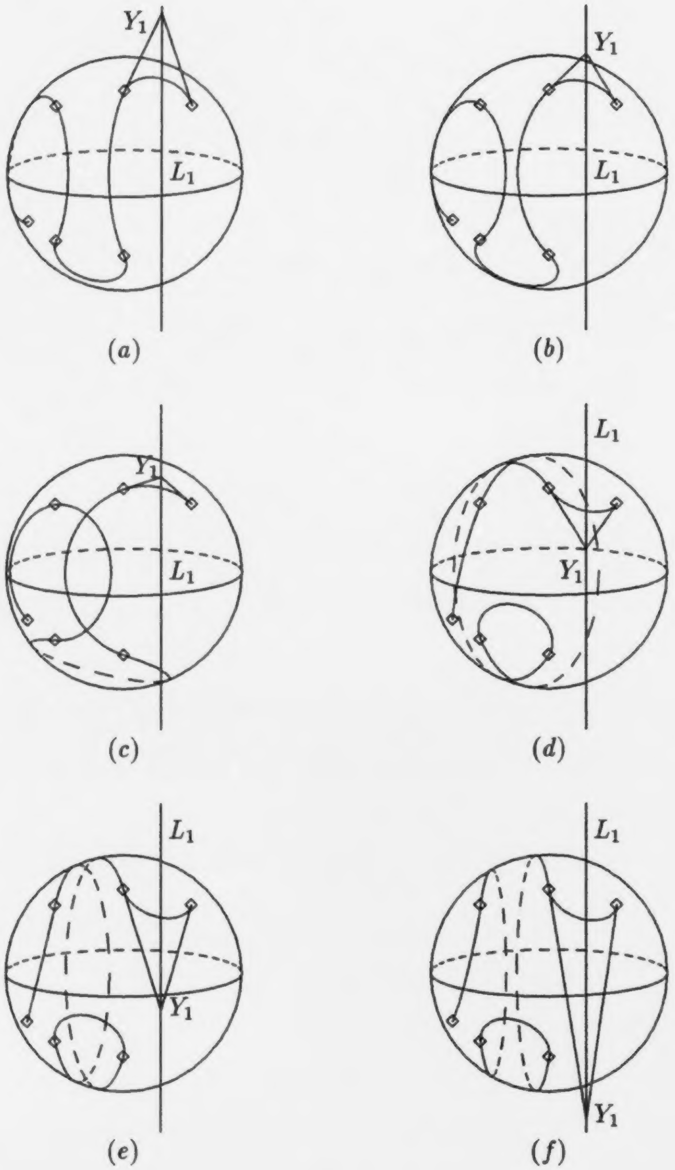

Y_1
L_1
(a)
Y_1
L_1
(b)
Y_1
L_1
(c)
L_1
Y_1
(d)
L_1
Y_1
(e)
L_1
Y_1
(f)

Figure 4.2.4 Interpolations on a sphere: Six different choices of the point $Y_1 \in L_1$ determine six different spline curves on a sphere interpolating the same set of data points marked by $\diamond$.

does not always have real fixed points on L_1.

When $d = 4$, L_1 is a 2-dimensional plane in E^4.

Lemma 4.2.13: *When* $d = 4$, M_{n+1} *always has a fixed point on the real projective plane* L_1.

PROOF: First establish a projective frame of reference F in L_1. Then the transformation induced by M_{n+1} on L_1 can be represented, with respect to F, by a nonsingular 3×3 real matrix M. Since any such matrix has a nonzero real eigenvalue and an associated real eigenvector, the projective transformation M_{n+1} has a real fixed point $Y_1 \in L_1$. □

Thus, in particular, for the closed interpolation problem on $S^3 \subset E^4$ we have

Theorem 4.2.14: *Let* $\{X_i\}_{i=1}^{n}$ *be a point sequence on the sphere* $S^3 \subset E^4$. *There exists a closed smooth rational quadratic spline interpolating* $\{X_i\}_{i=1}^{n}$ *on* S^3.

PROOF: By Lemma 4.2.13, $Y_1 \in L_1$ can be chosen to be a real fixed point of M_{n+1}. Let $\{X'_i\}_{i=1}^{n+2}$ be defined as $X'_i = X_i$, $i = 1, \ldots, n$, and $X'_{n+1} = X_1$, $X'_{n+2} = X_2$. Then by Theorem 4.2.11, there is a smooth rational quadratic spline interpolating $\{X'_i\}_{i=1}^{n+2}$ with the initial point as Y_1, which, by Theorem 4.2.12, is a closed spline interpolating $\{X_i\}_{i=1}^{n}$. □

For a closed interpolation problem on $S^3 \subset E^4$, the matrix M in Lemma 4.2.13 need not be constructed when a fixed point Y_1 of M_{n+1} in L_1 is wanted. One can instead find the real eigenvectors of M_{n+1} directly, one of which is guaranteed to be in L_1 by Lemma 4.2.13. The closed spline can then be computed using this point as Y_1.

4.3 Biarc interpolation on a quadric

In this section we consider the biarc interpolation problem on a quadric surface S formulated as follows. Let X_0, X_1 be two distinct points in normalized homogeneous form on S: $X^T AX = 0 \subset E^d$, $d \geq 3$, and assume that X_0, X_1 are on the same component but not on the same generating line of S. Let T_0, T_1 be the tangent directions to be interpolated at X_0, X_1, respectively. The problem is to find a biarc on S interpolating the data X_0, T_0, X_1 and T_1. Naturally we assume that T_0, T_1 are also tangent to S. Hence we have $X_0^T AT_0 = 0$, $X_1^T AT_1 = 0$.

A biarc on S is a composite curve consisting of two smoothly joining rational quadratic Bézier curves, or smooth conic arcs. The point where the two arcs join is called the *joint*

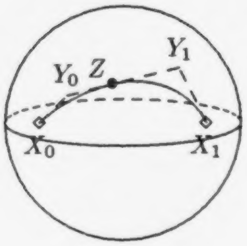


Figure 4.3.1 A spherical biarc: The data points are marked by ◇, and the joint point marked by •.

of the biarc. Since, in general, the complementary arc of a continuous conic arc is not continuous unless the given arc is elliptic, we will mainly consider only those special biarcs that are composed of rational quadratic curves with positive weights. These curves will be called *biarcs with positive weights.* First we shall establish a condition for the existence of a biarc with positive weights. Then the existence of the general biarc will be addressed. Finally several properties of the biarc will be derived.

4.3.1 Biarcs with positive weights

Our goal is to find a conic biarc B consisting of two conic arcs C_0, C_1 on S with standard Bézier representations $P_0(u)$, $P_1(v)$, $0 \leq u, v \leq 1$, such that tangent directions T_0, T_1 are interpolated at $X_0 = P_0(0)$ and $X_1 = P_1(1)$, and $P_0(u)$ and $P_1(v)$ share the same tangent direction at their joint Z and have positive weights.

Denote the three tangent hyperplanes of S at X_0, X_1 and Z by, respectively, Q_0: $X_0^T AX = 0$, Q_1: $X_1^T AX = 0$ and Q: $Z^T AX = 0$. Let the control points of $P_0(u)$ and $P_1(v)$ be X_0, Y_0, Z and Z, Y_1, X_1. By the discussion in the last section, Y_0 must be on the $(d-2)$-dimensional manifold $L_0 \equiv Q_0 \bigcap Q$: $X_0^T AX = Z^T AX = 0$. Similarly $Y_1 \in L_1 \equiv Q \bigcap Q_1$: $Z^T AX = X_1^T AX = 0$. In order that C_0, C_1 join smoothly at Z, the points Y_0, Z and Y_1 must be collinear. Let

$$\begin{aligned} Y_0 &= X_0 + k_0 T_0, \\ Y_1 &= X_1 - k_1 T_1, \end{aligned} \tag{4.6}$$

where $k_0, k_1 > 0$ and are yet to be determined. See Fig. 4.3.1. That $k_0, k_1 > 0$ follows from the assumption that only the biarcs with positive weights are being discussed. Consequently, by Lemma 4.1.1, the joint Z is between Y_0 and Y_1. Now we assume that $[Y_0] \neq [Y_1]$; therefore the straight line Y_0Y_1 is uniquely defined.

Since $X_0^T AX_0 = X_0^T AT_0 = X_1^T AX_1 = X_1^T AT_1 = 0$, it follows from (4.6) that

$$\begin{aligned} Y_0^T AY_0 &= k_0^2 T_0^T AT_0, \\ Y_1^T AY_1 &= k_1^2 T_1^T AT_1. \end{aligned} \tag{4.7}$$

As a solution to the above biarc interpolation problem can be regarded as a solution to the point interpolation problem discussed in the last section for the data points X_0, Z and X_1, we obtain the following necessary condition.

Lemma 4.3.1: *A necessary condition for the biarc interpolation problem to have a solution is*

$$(T_0^T AT_0)(T_1^T AT_1) > 0.$$

PROOF: When the problem is solvable, by Lemma 4.2.10, $(Y_0^T AY_0)(Y_1^T AY_1) > 0$. By (4.7), since $k_0^2 k_1^2 > 0$, we obtain $(T_0^T AT_0)(T_1^T AT_1) > 0$. □

We now proceed to derive an equation that governs k_0 and k_1. Because of Lemma 4.3.1, without loss of generality, we can normalize T_0, T_1, replacing A by $-A$ if necessary, such that $T_0^T AT_0 = T_1^T AT_1 = 1$. From now on we will always assume that T_0 and T_1 are so normalized. Thus (4.7) can be written as

$$\begin{aligned} Y_0^T AY_0 &= k_0^2, \\ Y_1^T AY_1 &= k_1^2. \end{aligned} \tag{4.8}$$

By the preceding observation regarding the relation between Y_0, Z and Y_1, the straight line Y_0Y_1 is well defined and Z is the tangent point of the line $Y_0Y_1 : \lambda Y_0 + \mu Y_1$ to S. Thus the Joachimsthal's equation [SeK52], obtained by substituting the point $\lambda Y_0 + \mu Y_1$ in the equation of S,

$$\lambda^2(Y_0^T AY_0) + 2\lambda\mu(Y_0^T AY_1) + \mu^2(Y_1^T AY_1) = 0,$$

has a double root. Therefore the discriminant

$$\Delta \equiv 4[(Y_0^T AY_1)^2 - (Y_0^T AY_0)(Y_1^T AY_1)] = 0,$$

or, by (4.8)

$$(Y_0^T AY_1)^2 - k_0^2 k_1^2 = 0.$$

Then we have

$$Y_0^T AY_1 - k_0 k_1 = 0 \tag{4.9}$$

or

$$Y_0^T A Y_1 + k_0 k_1 = 0. \tag{4.10}$$

When $\Delta = 0$, $\lambda/\mu = -(Y_0^T A Y_1)/(Y_0^T A Y_0)$. Thus, omitting a nonzero multiplicative factor, the straight line $\lambda Y_0 + \mu Y_1$ touches S at

$$Z = (Y_0^T A Y_1) Y_0 - (Y_0^T A Y_0) Y_1.$$

By (4.9) or (4.10) we obtain, respectively,

$$Z = k_0 k_1 Y_0 - k_0^2 Y_1 \tag{4.11}$$

or

$$Z = -k_0 k_1 Y_0 - k_0^2 Y_1. \tag{4.12}$$

Since Y_0 and Y_1 are in normalized homogeneous form and Z is required to lie between Y_0 and Y_1, we discard (4.11) and retain (4.12) as the desired expression for Z, because when $k_0, k_1 > 0$, (4.11) gives Z outside line segment $\overline{Y_0 Y_1}$. Dividing by $-k_0$ in (4.12) yields

$$\begin{aligned} Z(k_0, k_1) = k_1 Y_0 + k_0 Y_1 &= k_1(X_0 + k_0 T_0) + k_0(X_1 - k_1 T_1) \\ &= k_1[X_0 + k_0(T_0 - T_1)] + k_0 X_1. \end{aligned} \tag{4.13}$$

Substituting (4.6) in (4.10), k_0 and k_1 are related by

$$X_0^T A X_1 + k_0 X_1^T A T_0 - k_1 X_0^T A T_1 + k_0 k_1 (1 - T_0^T A T_1) = 0. \tag{4.14}$$

In the above derivation it has been assumed that $[Y_0] \neq [Y_1]$; for otherwise the straight line $Y_0 Y_1$ is not uniquely defined. It will be shown that $[Y_0] = [Y_1]$ for some k_0,k_1 satisfying (4.14) occurs only for a class of data with a special configuration.

Definition 4.3.1: *The data $D = \{X_0, T_0, X_1, T_1\}$ is* singular *if $X_0 + \rho T_0 = X_1 + \rho T_1$ for some finite $\rho \neq 0$ or $T_0 = T_1$. The data that is not singular is called* regular.

Note that if D is singular then the four points X_0, T_0, X_1, T_1 are coplanar, i.e. linearly dependent; here T_0 and T_1 are regarded as points at infinity. But the converse is not true. An example of singular data is illustrated in Fig. 4.3.2.

Lemma 4.3.2: *Given data $D = \{X_0, T_0, X_1, T_1\}$ on the quadric surface S, $[Y_0] = [Y_1]$ for some k_0 and k_1 satisfying (4.14) if and only if D is singular, where Y_0 and Y_1 are given by (4.6) and k_0, k_1 may be of any sign.*

PROOF: First consider necessity. There are two cases to consider: (i) $[Y_0] = [Y_1]$ is a finite point; (ii) $[Y_0] = [Y_1]$ is a point at infinity.

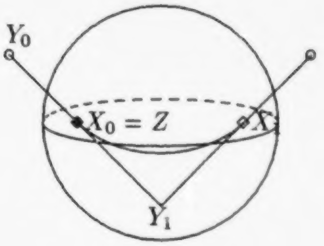


Figure 4.3.2 Singular data: An instance of singular data on S^2 is shown with one of its degenerate biarc interpolants and the control polygons. The joint point Z is marked with •, which coincides with X_0. The ends of tangents T_0 and T_1 are marked with ∘.

(i) In this case k_0 and k_1 are finite and $[Y_0] = [Y_1]$ implies that $Y_0 = Y_1$. Since X_0, X_1 are distinct points, $k_0 \neq 0$ or $k_1 \neq 0$; for otherwise from $Y_0 = Y_1$ and (4.6), $X_0 = X_1$ would result. First assume $k_0 \neq 0$. By (4.6),

$$Y_0^T A Y_0 = (X_0 + k_0 T_0)^T A (X_0 + k_0 T_0) = k_0^2.$$

On the other hand, since $Y_0 = Y_1$ and k_0,k_1 satisfy (4.10), which is equivalent to (4.14), $Y_0^T A Y_0 = Y_0^T A Y_1 = -k_0 k_1$. Therefore $k_0^2 = -k_0 k_1$, or $k_0 = -k_1$ since $k_0 \neq 0$. Thus from $Y_0 = Y_1$, we obtain

$$X_0 + k_0 T_0 = X_1 - k_1 T_1 = X_1 + k_0 T_1.$$

Hence, by definition, D is singular. When $k_1 \neq 0$, the same conclusion follows from a symmetric argument.

(ii) In this case k_0 and k_1 are infinite. From $[Y_0] = [Y_1]$, we have either $T_0 = T_1$ or $T_0 = -T_1$. Eqn. (4.14) can be rewritten as

$$\frac{X_0^T A X_1}{k_0 k_1} + \frac{X_1^T A T_0}{k_1} - \frac{X_0^T A T_1}{k_0} + 1 - T_0^T A T_1 = 0,$$

which is satisfied by k_0, k_1 when $T_0 = T_1$ but not when $T_0 = -T_1$, since when $T_0 = T_1$, $1 - T_0^T A T_1 = 1 - T_0^T A T_0 = 0$, but when $T_0 = -T_1$, $1 - T_0^T A T_1 = 1 + T_0^T A T_0 = 2$. Thus we are left with $T_0 = T_1$. Hence D is singular.

Now we prove sufficiency. Suppose that D is singular. When $X_0 + \rho T_0 = X_1 + \rho T_1$ for some finite $\rho \neq 0$, then using this equality, it can be verified that $k_0 = \rho$ and $k_1 = -\rho$ satisfy (4.14). For this pair of k_0 and k_1, $Y_0 = Y_1$. When $T_0 = T_1$, as above it can be shown again that $k_0 = \pm\infty$ and $k_1 = \pm\infty$ satisfy (4.14). For this pair of k_0 and k_1, we have $[Y_0] = [Y_1] = [T_0] = [T_1]$. □

Another necessary condition for the biarc interpolation problem to have a biarc with positive weights as a solution is that (4.14) has solutions $k_0 > 0$ and $k_1 > 0$. From the derivation of the above conditions on k_0 and k_1, we have obtained the following necessary and sufficient condition on the existence of a biarc with positive weights when the data is regular, which can also be used to construct the biarc if it exists.

Theorem 4.3.3: *There exists a biarc with positive weights interpolating the regular data* $D = \{X_0, T_0, X_1, T_1\}$ *on the quadric surface* S*:* $X^TAX = 0 \;\subset E^d$ *if and only if* $(T_0^TAT_0)(T_1^TAT_1) > 0$*, and Eqn. (4.14) has solutions* $k_0 > 0$ *and* $k_1 > 0$*, assuming that* T_0 *and* T_1 *have been normalized to* $T_0^TAT_0 = T_1^TAT_1 = 1$*, replacing* A *by* $-A$ *if necessary.*

PROOF: The necessity has been shown in the above discussion. For the sufficiency we observe that when the conditions hold a joint Z between Y_0 and Y_1 is given by (4.13), and then by Lemma 4.1.1, a biarc with positive weights can be constructed.□

While the first condition of Theorem 4.3.3 is easy to and understand, the second condition can only serve in algorithmic treatments. Regarding (4.14) as a hyperbola in k_0-k_1 plane, the conditons for the existence of solutions $k_0 > 0$ and $k_1 > 0$ can be readily established, which is equivalent to that the hyperbola passes through the first quadrant. But the geometric significance of the second condition is yet to be understood. That is, for a general quadric surface S, we still do not know the geometric characteristic of a data set which satisfies the second condition, or how restrictive this condition really is. However, on the sphere $S^{d-1} \subset E^d$ this condition is satisfied for generic data.

Before giving the result on S^{d-1} we need to distinguish a degenerate biarc solution from a proper biarc solution to the biarc interpolation problem. There are some cases where the biarc interpolating the data D becomes degenerate.

Definition 4.3.2: *A biarc is* degenerate *if one of its arcs reduces to a single point. A biarc that is not degenerate is called* proper.

A biarc with control points X_0, Y_0, Z and Z, Y_1, X_1 is degenerate if and only if Z coincides with X_0 or X_1. The necessity is obvious. For the sufficiency suppose that $[Z] = [X_1]$; the other case is similar. Then the control polygon ZY_1X_1 collapses into two coincident line segments. First assume $k_1 \neq 0$. Since $Z^TAX_1 = 0$ and $Y_1^TAY_1 = k_1^2 \neq 0$, the weight of the second arc is 0 by (4.4), i.e. the arc controlled by ZY_1X_1 is a point. When $k_1 = 0$, by (4.6), $[Z] = [Y_1] = [X_1]$, and again the arc becomes a point.

In the following both proper and degenerate biarcs are regarded as solutions to the biarc interpolation problem since in both cases Eqn.(4.14) is satisfied by nonzero k_0 and k_1. But we are only interested in finding the proper biarc with positive weights. Next we shall show that for regular data on S^{d-1}, which is generic, a proper biarc with positive weight always exists.

The following lemmas assert that given data D on S, a biarc with positive weights and $k_0k_1 \neq 0$ is degenerate if and only if D is singular. Presently we avoid discussing the case $[Y_0] = [Y_1]$, since then the argument leading to (4.14) is not valid. This case will be dealt with in Section 4.3.2.

Lemma 4.3.4: *For singular data D on a quadric surface S, all the biarcs with positive weights and $k_0 \neq 0$, $k_1 \neq 0$ are degenerate, where k_0,k_1 satisfy (4.14) and are such that $[Y_0] \neq [Y_1]$.*

PROOF: Let $D = \{X_0, T_0, X_1, T_1\}$ be singular. There are two cases to consider: (i) $X_0 + \rho T_0 = X_1 + \rho T_1$ for some finite $\rho \neq 0$; (ii) $T_0 = T_1$.

(i) In this case $T_0 \neq T_1$. The left hand side of Eqn.(4.14) becomes

$$\begin{aligned}
& X_0^T A(X_0 + \rho T_0 - \rho T_1) + k_0(X_0 + \rho T_0 - \rho T_1)^T A T_0 - k_1(X_0^T A T_1) + k_0 k_1(1 - T_0^T A T_1) \\
= \; & -\rho(X_0^T A T_1) + k_0\rho(1 - T_0^T A T_1) - k_1(X_0^T A T_1) + k_0 k_1(1 - T_0^T A T_1) \\
= \; & (k_1 + \rho)[k_0(1 - T_0^T A T_1) - (X_0^T A T_1)] \\
= \; & (k_1 + \rho)[k_0(1 - T_0^T A T_1) - (X_1 + \rho T_1 - \rho T_0)^T A T_1] \\
= \; & (k_1 + \rho)[k_0(1 - T_0^T A T_1) - \rho(1 - T_0^T A T_1)] \\
= \; & (1 - T_0^T A T_1)(k_0 - \rho)(k_1 + \rho).
\end{aligned}$$

Since the left side of Eqn.(4.14) is not identically zero as $X_0^T A X_1 \neq 0$, it factors, with solutions $k_0 = \rho$ or $k_1 = -\rho$. First we take $(k_0, k_1) = (\rho, k_1)$, $k_1 \neq -\rho$, as solutions of (4.14). Here $k_1 \neq -\rho$ since $k_0 = \rho$ and $k_1 = -\rho$ would cause $Y_0 = Y_1$. Therefore by (4.13),

$$Z = k_1(X_0 + k_0 T_0) + k_0(X_1 - k_1 T_1) = k_1(X_0 + \rho T_0 - \rho T_1) + \rho X_1 = (k_1 + \rho) X_1.$$

Thus $[Z] = [X_1]$ since $k_1 + \rho \neq 0$. Hence the resulting biarc is degenerate. When $(k_0, -\rho)$, where $k_0 \neq \rho$, are taken as solutions of (4.14), it can be similarly shown that $[Z] = [X_0]$.

(ii) $T_0 = T_1$. In this case $1 - T_0^T A T_1 = 1 - T_0^T A T_0 = 0$, $X_0^T A T_1 = X_0^T A T_0 = 0$, and $X_1^T A T_0 = X_1^T A T_1 = 0$. Therefore (4.14) can be rewritten as

$$\frac{X_0^T A X_1}{k_0 k_1} + \frac{X_1^T A T_0}{k_1} - \frac{X_0^T A T_1}{k_0} + 1 - T_0^T A T_1 = 0,$$

which reduces to $(X_0^T A X_1)/(k_0 k_1) = 0$. This equation is satisfied by $k_0 = \pm\infty$ or $k_1 = \pm\infty$. But when $k_0 = \pm\infty$ and $k_1 = \pm\infty$, $[Y_0] = [Y_1] = [T_0]$, which is excluded by the assumption of the theorem. Therefore one of k_0 and k_1 is finite. First take a finite k_0 and $k_1 = \infty$ as solutions of (4.14). Then

$$\begin{aligned} [Z] &= [k_1(X_0 + k_0T_0) + k_0(X_1 - k_1T_1)] = [k_1(X_0 + k_0T_0 - k_0T_1) + k_0X_1] \\ &= [k_1X_0 + k_0X_1] = [X_0]. \end{aligned}$$

Hence the resulting biarc is degenerate. The cases in which k_0 and k_1 take other combinations of values can be proved similarly. □

Lemma 4.3.5: *For regular data D on a quadric surface S, all its biarcs with $k_0 > 0$ and $k_1 > 0$ are proper, where k_0 and k_1 satisfy (4.14).*

PROOF: We just need to show that any solution k_0, k_1 of (4.14), where k_0, $k_1 > 0$, results in a proper biarc. Suppose the opposite. Then for regular $D = \{X_0, T_0, X_1, T_1\}$ there is a degenerate biarc with k_0, $k_1 > 0$. Without loss of generality, assume that $Z = \beta X_0$ for some $\beta \neq 0$; then by (4.13),

$$\beta X_0 = k_1(X_0 + k_0T_0) + k_0(X_1 - k_1T_1).$$

Since X_0 and X_1 are in normalized homogeneous form and the last components of T_0 and T_1 are zero, $\beta = k_0 + k_1$. Thus

$$k_0X_0 - k_0k_1T_0 = k_0X_1 - k_0k_1T_1,$$

or, since $k_0 \neq 0$,

$$X_0 - k_1T_0 = X_1 - k_1T_1.$$

Since $k_1 \neq 0$, the data D is singular. This is a contradiction. □

Theorem 4.3.6: *There exists a proper biarc with positive weights for data D on $S^{d-1} \subset E^d$ if and only if D is regular.*

PROOF: The necessity is implied by Lemma 4.3.4. We only have to prove the sufficiency. By Lemma 4.3.5, since D is regular, every biarc for D with k_0, $k_1 > 0$ must be proper. It suffices to show that the two conditions (i) and (ii) of Theorem 4.3.3 are always met on S^{d-1}. Then it will follow that a proper biarc with positive weights for D exists.

Let the equation of S^{d-1} be $X^TAX = 0$, with

$$A = \begin{bmatrix} I_d & 0 \\ 0 & -1 \end{bmatrix},$$

where I_d is the $d \times d$ identity matrix. Since $T_0^T A T_0 = T_0^T T_0 > 0$ and $T_1^T A T_1 = T_1^T T_1 > 0$, obviously, the first condition of Theorem 4.3.3 is always satisfied. The second condition is that Eqn.(4.14), or in this case

$$X_0^T A X_1 + k_0 X_1^T T_0 - k_1 X_0^T T_1 + k_0 k_1 (1 - T_0^T T_1) = 0, \tag{4.15}$$

has positive k_0, k_1 as solutions. Since X_0, X_1 are in normalized homogeneous form, and $X_0 \neq X_1$, we have

$$\begin{aligned} (X_0 - X_1)^T (X_0 - X_1) &= (X_0 - X_1)^T A (X_0 - X_1) \\ &= -2 X_0^T A X_1 > 0. \end{aligned}$$

Thus $X_0^T A X_1 < 0$, i.e. the constant term of (4.15) is negative. Since D is regular, by definition, $T_0 \neq T_1$. Therefore

$$1 - T_0^T T_1 = \frac{1}{2} (T_0 - T_1)^T (T_0 - T_1) > 0. \tag{4.16}$$

That is, the coefficient of $k_0 k_1$ in (4.15) is positive. Hence (4.15) has positive solutions k_0 and k_1 because $k_0 = k_1 = 0$ makes the left hand side of (4.15) negative and sufficiently large positive values of k_0, k_1 make it positive. □

Setting $k_0 = k_1$ in (4.14) yields the equation

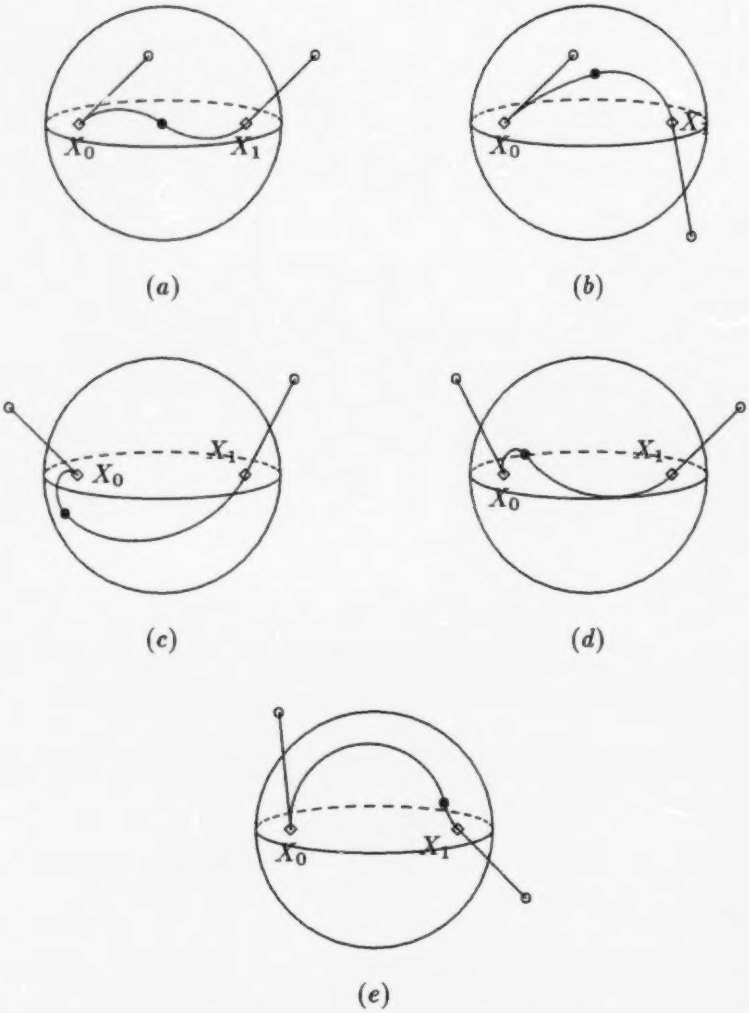
$$ak^2 + bk + c = 0, \tag{4.17}$$

where $a = 1 - T_0^T A T_1$, $b = X_1^T A T_0 - X_0^T A T_1$ and $c = X_0^T A X_1$. As argued in the above proof, this equation has positive solution $k = [-b + (b^2 - 4ac)^{1/2}]/(2a)$ for regular data D on S^{d-1}. Therefore a particular positive solution of (4.14), in this case, is $k_0 = k_1 = k$. Fig. 4.3.3 shows the biarc interpolants on S^2 for several different data configurations, using the positive root of (4.17) as k_0 and k_1.

4.3.2 General biarcs

In the previous subsection we concentrated on the existence of biarcs with positive weights. Corresponding but weaker results exist for general biarcs, since in general a rational quadratic curve with a negative weight is not continuous. In this subsection we will examine the consequence of allowing negative weights in a biarc when the resulting curve is continuous. Such a biarc is called a *general biarc*.

If negative weights are allowed, then the parameters k_0 and k_1 are not restricted to be positive. Let a biarc interpolating $D = \{X_0, T_0, X_1, T_1\}$ consist of two rational quadratic



(a) (b)

(c) (d)

(e)

Figure 4.3.3 Biarc interpolations on a sphere: Five different data configurations on a sphere and their biarc interpolants. The data points X_0 and X_1 are marked with $\diamond$, the ends of T_0 and T_1 with $\circ$, and the joint point Z with $\bullet$. Here the parameters k_0 and k_1 of each interpolant are the positive root of Eqn. (4.17).

Bézier curves C_0 and C_1 with weights w_0 and w_1. Let the control points of C_0 and C_1 be X_0, Y_0, Z and Z, Y_1, X_1, where Y_0 and Y_1 are given by (4.6). A simple analysis shows that, to meet the end tangent conditions, the weight w_i should be positive if $k_i > 0$, and negative if $k_i < 0$, $i = 0, 1$. Then it can be shown, with the aid of Lemma 4.1.1, that by choosing such weights according to the nonzero solutions k_0 and k_1 of (4.14), the two Bézier curves meet smoothly at Z, but the continuity of each curve is not guaranteed unless the weight is positive. Also it can be shown that by choosing such weights to meet the end tangent conditions, the two resulting curves corresponding to any nonzero solutions k_0 and k_1 of (4.9) can not meet smoothly because the tangents of the two arcs at Z point opposite directions. This means that we do not have to consider the solutions of (4.9) even when studying general biarcs. Note that it is for a different reason that (4.9) was rejected in the previous discussion of the biarc with positive weights.

When the general biarc is allowed, singular data may also have proper biarc solutions. By Lemma 4.3.4, for singular data D, when $Y_0 \neq Y_1$ the locus of the joint Z given by (4.13) consists of two isolated points X_0 and X_1; in this case the resulting biarcs are easily seen to be degenerate. When $Y_0 = Y_1$ for some k_0, k_1, the arguments leading to (4.14) break down since the straight line Y_0Y_1 is not uniquely defined; Lemma 4.3.2 shows that this happens if and only if D is singular.

When $Y_0 = Y_1$ for singular data $D = \{X_0, T_0, X_1, T_1\}$, where $X_0 + \rho T_0 = X_1 + \rho T_1$ for some $\rho \neq 0$, we have $Y_0 = X_0 + \rho T_0$ and $Y_1 = X_1 + \rho T_1$. Therefore $k_0 = \rho$ and $k_1 = -\rho$. Because Z is only required to be the tangent point to S of a straight line passing through $Y_0 = Y_1$, the locus of Z is the intersection of S with the polar hyperplane $Y_0^T AX = 0$ of Y_0, which is denoted by J. For each point $Z \in J$ but $Z \neq X_0, X_1$, when the point $Y_0 = Y_1$ is finite $k_0 k_1 = -\rho^2 < 0$; therefore one of the two resulting Bézier curves has the negative weight; when $Y_0 = Y_1$ is at infinity, i.e. in the case of $T_0 = T_1$, the two curves are semi-ellipses. These two Bézier curves yield a proper biarc interpolating D if they are both continuous. See Fig. 4.3.4 for an illustration of two general biarcs on S^2 for singular data in which $T_0 = T_1$. For singular data D, if proper biarcs exist then their degrees of freedom, which are the degrees of freedom of $Z \in J$, are $d - 2$. We will see later on that the degree of freedom of proper biarcs for regular data, if they exist, is always one. Therefore, when $d > 3$, it is probable that more proper biarcs exist for singular data than for regular data. This is exactly the case on the sphere $S^{d-1} \subset E^d$, $d > 3$.

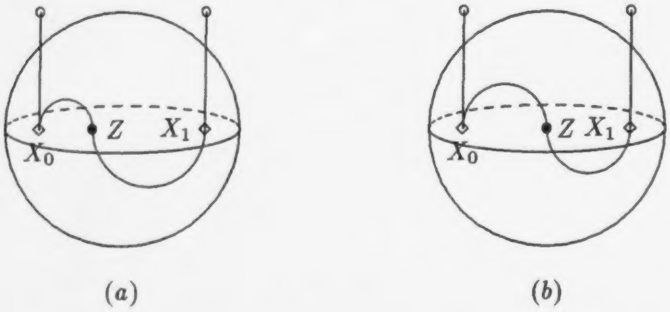


Figure 4.3.4 Biarc interpolatons for singular data: On S^2 when $T_0 = T_1$ the data is singular. There is a family of biarcs with the joint point Z, marked with •, lying on the great circle passing throuh X_0 and X_1.

4.3.3 Properties of biarcs with positive weights

In this subsection we prove two properties concerning the locus of the joint Z. When (4.14) does not factor into two linear terms, k_1 can be expressed in terms of k_0; then $Z(k_0, k_1)$ given by (4.13) will be a parametric curve in k_0.

Lemma 4.3.7: *Eqn. (4.14) does not factor into linear factors for coplanar regular data and generic noncoplanar regular data.*

PROOF: Since a bilinear function $axy + bx + cy + d$, with $a \neq 0$, factors if and only if the discriminant $ad - bc = 0$, we just need to show that the discriminant of the left hand side of (4.14) does not vanish for the stated data.

First consider generic noncoplanar regular data D. Let $\mathcal{F}$ be the 3-dimensional projective space spanned by the four points of D. We assume that D is generic in the sense that the 2-dimensional quadric $\bar{S}$ intersected by $\mathcal{F}$ from $S:\ X^TAX = 0$ is regular. Let h_1 and h_2 be the discriminants of (4.9) and (4.10) respectively, after substituting (4.6), i.e.

$$h_1 = -(X_0^TAX_1)(1 + T_0^TAT_1) + (X_0^TAT_1)(X_1^TAT_0),$$

and

$$h_2 = (X_0^TAX_1)(1 - T_0^TAT_1) + (X_0^TAT_1)(X_1^TAT_0).$$

Let $M = [X_0\ T_0\ X_1\ T_1]^TA[X_0\ T_0\ X_1\ T_1]$. It can be verified that $\det(M) = h_1h_2$. Since the quadric $\bar{S}$ is regular, $\det(M) \neq 0$ as M is the coefficient matrix of $\bar{S}$. Hence $h_2 \neq 0$.

When D is coplanar regular data, noting that the last components of T_0, T_1 are zero and X_0, X_1 are in normalized form, $X_0 + \rho_0 T_0 = X_1 + \rho_1 T_1$ for some ρ_0 and ρ_1. Then

$$(X_0 + \rho_0 T_0)^T A (X_0 + \rho_0 T_0) = (X_1 + \rho_1 T_1)^T A (X_1 + \rho_1 T_1),$$

or, after simplifying, $\rho_0^2 = \rho_1^2$. Since D is not singular, we conclude that $\rho_0 = -\rho_1$. Letting $\rho = \rho_0 = -\rho_1$, we have $X_0 + \rho T_0 = X_1 - \rho T_1$; obviously $\rho \neq 0$ since $X_0 \neq X_1$. Using this equality it can be verified directly that $h_2 = 2X_0^T A X_1 \neq 0$, since X_0,X_1 are not on the same generating line of S. Hence $h_2 \neq 0$ for coplanar regular data and generic noncoplanar regular data, i.e. Eqn. (4.14) does not factor for these data.□

Theorem 4.3.8: *For coplanar regular data and generic noncoplanar regular data* $D = \{X_0, T_0, X_1, T_1\}$ *on a quadric surface* S*, the locus of* Z *given by (4.13) is a conic on* S *passing through* X_0 *and* X_1.

PROOF: By Lemma 4.3.7, (4.14) is irreducible, so k_1 can be expressed in terms of k_0,

$$k_1 = \frac{X_0^T A X_1 + k_0 X_1^T A T_0}{X_0^T A T_1 + k_0 (T_0^T A T_1 - 1)}.$$

Substituting this expression in (4.13) and clearing the denominator, we have

$$Z(k_0) = [X_0^T A X_1 + k_0 X_1^T A T_0][X_0 + k_0(T_0 - T_1)] + k_0[X_0^T A T_1 + k_0(T_0^T A T_1 - 1)]X_1. \quad (4.18)$$

So the locus of Z is a rational quadratic curve on S. Since $Z = (X_0^T A X_1)X_0$ when $k_0 = 0$, the locus passes through X_0. Symmetrically, it passes through X_1.□

Theorem 4.3.9: *When the data* $D = \{X_0, T_0, X_1, T_1\}$ *are taken from a proper conic arc* C *on a quadric surface* S*, any proper biarc with positive weights interpolating* D *reproduces* C.

PROOF: Obviously, C provides a biarc solution to D. Now we show that the locus J of the joint Z for D coincides with the underlying conic of C. Then it will follow that all the biarcs with positive weights, which must have their joints on C, reproduce C.

By Theorem 4.3.8, $[Z(k_0)_{k_0=0}] = [X_0]$. And by (4.18),

$$\begin{aligned} Z'(k_0) &= (X_1^T A T_0)[X_0 + k_0(T_0 - T_1)] + [X_0^T A X_1 + k_0 X_1^T A T_0)](T_0 - T_1) \\ &\quad + [X_0^T A T_1 + 2k_0(T_0^T A T_1 - 1)]X_1 \end{aligned}$$

Then

$$\begin{aligned} Z'(k_0)|_{k_0=0} &= (X_1^T A T_0)X_0 + (X_0^T A X_1)(T_0 - T_1) + (X_0^T A T_1)X_1 \\ &= (X_1^T A T_0)X_0 + (X_0^T A X_1)T_0 - (X_0^T A X_1)T_1 + (X_0^T A T_1)X_1. \end{aligned}$$

Since D is coplanar but not singular, as has been shown in the above proof of Lemma 4.3.7, $X_0 + \rho T_0 = X_1 - \rho T_1$ for some $\rho \neq 0$, or

$$X_1 = X_0 + \rho T_0 + \rho T_1.$$

So

$$\begin{aligned}
Z'(k_0)|_{k_0=0} &= (X_1^T A T_0)X_0 + (X_0^T A X_1)T_0 \\
&\quad -[X_0^T A(X_0 + \rho T_0 + \rho T_1)]T_1 + (X_0^T A T_1)(X_0 + \rho T_0 + \rho T_1) \\
&= (X_1^T A T_0)X_0 + (X_0^T A X_1)T_0 - \rho(X_0^T A T_1)T_1 \\
&\quad +(X_0^T A T_1)X_0 + \rho(X_0^T A T_1)T_0 + \rho(X_0^T A T_1)T_1 \\
&= (X_1^T A T_0 + X_0^T A T_1)X_0 + (X_0^T A X_1 + \rho X_0^T A T_1)T_0.
\end{aligned}$$

Therefore $Z'(k_0)|_{k_0=0}$ is on the straight line X_0T_0, i.e. $Z(k_0)$ shares the same tangent with C at X_0. By symmetry, the same holds at X_1. Hence the locus of Z coincides with the underlying conic of C, since they are both the conic obtained by intersecting S with the 2-dimensional plane containing C. □

4.4 Point interpolation using biarcs

In Section 4.2 we saw that when a rational quadratic spline is used to interpolate a point sequence $\{X_i\}_{i=1}^n$ on a quadric surface, if the solution exists it is determined by a global parameter $Y_1 \in L_1$, which has $(d-2)$ degrees of freedom. This property is quite undesirable because any local perturbation of the data or change of Y_1 would have global influence on the curve. Now we consider using the biarc interpolant to obtain a locally controllable interpolating spline to $\{X_i\}_{i=1}^n$. Of course, more curve segments are needed to achieve this goal than in the previous approach.

The algorithm is given as follows.

Algorithm 4.4.1:

Given $\{X_i\}_{i=1}^n$ on the same component of a regular quadric surface $S \subset E^d$, $d \geq 3$, such that no two consecutive points X_i, X_{i+1} are on the same generating line of S, and $(X_1^T A X_2)(X_i^T A X_{i+1}) > 0$, $i = 1, 2, \ldots, n-1$.

1. Determine the tangent T_i associated with X_i as the tangent to the conic $\tilde{C}_i$ on S interpolating the three points X_{i-1}, X_i and X_{i+1}, $i = 1, 2, \ldots, n$. The direction of

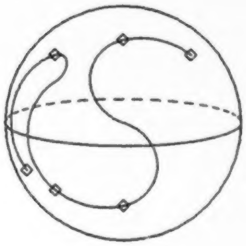


Figure 4.4.1 Spline interpolation using biarcs: The same data points as in Fig. 4.2.4 are interpolated using the biarc interpolant as described in Algorithm 4.4.1. The parameters k_0, k_1 of each biarc interpolant are given by the positive root of Eqn. (4.17).

each tangent conforms with the direction along which a variable point travels on $\tilde{C}_i$ from X_{i-1} through X_i to X_{i+1} without covering the whole $\tilde{C}_i$. Here we assume that $X_0 = X_3$ and $X_{n+1} = X_{n-2}$ to provide the end tangent directions.

2. Use a biarc with positive weights to interpolate $D_i = \{X_i, T_i, X_{i+1}, T_{i+1}\}, i = 1, \ldots, n-1$. □.

The tangent T_i in the above algorithm can be computed as follows. Let $T_i = [t_1, t_2, \ldots, t_{d+1}]^T$. Since T_i is in the plane determined by X_{i-1},X_i and X_{i+1}, we can write $T_i = aX_{i-1} + bX_i + cX_{i+1}$. The constants a, b and c can be uniquely determined by the following constraints: (i) $t_{d+1} = 0$, which means that T_i is a point at infinity; (ii) $X_i^T A T_i = 0$, i.e. T_i is on the tangent hyperplane of S at X_i; (iii) $(X_{i+1} - X_{i-1})^T T_i > 0$, which determines the direction on $\tilde{C}_i$ at X_i from X_{i-1} to X_{i+1}; and (iv) $T_i^T A T_i = 1$, which is required by the algorithm.

Fig. 4.4.1 shows the effect of using Algorithm 4.4.1 to interpolate the same data points as in the Fig. 4.2.4 by biarcs. The parameters of each biarc are determined by (4.17), i.e. $k_0 = k_1 > 0$.

A feature of the above algorithm is that conic sections can be locally reproduced in the sense that if X_{i-1}, X_i, X_{i+1} and X_{i+2} are on any conic C on S, then C is reproduced between X_i and X_{i+1} by the algorithm, $i = 1, \ldots, n-1$. This property follows from the scheme of assigning tangents to $\{X_i\}_{i=1}^n$ described in the above algorithm and Theorem 4.3.9.

According to the results established so far we can only be sure that the above algorithm works correctly for a point sequence on a surface S which is affinely equivalent to S^{d-1}. For a general quadric surface the restriction $(X_1^T AX_2)(X_i^T AX_{i+1}) > 0$ on the input $\{X_i\}_{i=1}^n$ is necessary but not sufficient for this biarc scheme. For if a smooth interpolating spline exists, and if Z_i is the joint of the biarc between X_i and X_{i+1}, then, by Theorem 4.2.9, all the line segments $\overline{X_iZ_i}$ and $\overline{Z_iX_{i+1}}$, $i = 1, \ldots, n-1$, are on the same side of S. But for any i, $\overline{X_iZ_i}$ and $\overline{Z_iX_{i+1}}$ are on the same side of S as the segment $\overline{X_iX_{i+1}}$; therefore all $\overline{X_iX_{i+1}}$, $i = 1, \ldots, n-1$, are on the same side of S, i.e. $(X_1^T AX_2)(X_i^T AX_{i+1}) > 0$.

In the second step of Algorithm 4.4.1, if there exist biarcs interpolating D_i, we have to choose one of them according to some criterion. A satisfactory choice would entail a study of the influence of the parameters k_0 and k_1 on the shape of the resulting biarc. On a general quadric this problem remains open. In the case of a sphere this problem will be addressed in the next two chapters.

4.5 Summary

We have studied two problems: point interpolation and biarc interpolation on a regular quadric surface $S \subset E^d$, $d \geq 3$. Given a point sequence $\{X_i\}_{i=1}^n$, $n \geq 3$, on a real component of $S : X^T AX = 0$, it is shown that a necessary condition for a nontrivial rational quadratic spline on S interpolating $\{X_i\}_{i=1}^n$ to exist is $(X_1^T AX_2)(X_i^T AX_{i+1}) > 0$, $i = 1, 2, \ldots, n-1$, or geometrically, all the line segments $\overline{X_iX_{i+1}}$, $i = 1, 2, \ldots, n-1$, are on the same side of S. This condition is sufficient and always true when S is affinely equivalent to a sphere.

For the second problem, let X_0, X_1 be distinct points on S and T_0, T_1 be tangent directions to be interpolated at X_0 and X_1, respectively. The interpolant we use, called the biarc interpolant, is composed of two smoothly joining rational quadratic segments. It is shown that for generic data this scheme provides a solution with one degree of freedom when the solution exists. A necessary and sufficient condition for the existence of a biarc with positive weights is given. It is shown that this condition is satisfied by generic data on the sphere $S^{d-1} \subset E^d$, $d \geq 3$.

Several open problems still remain. We have shown that not every point sequence on a general quadric surface admits a rational quadratic spline interpolant. So as this kind of data occurs in practice, an alternate scheme should be developed to cope with it.

For the biarc interpolation problem we have given a necessary and sufficient condition for the existence of the interpolant. Yet, except for the sphere S^{d-1}, where the condition is always fulfilled, we do not know how restrictive this condition really is on a general quadric surface in terms of a geometric characterization. For instance, how restrictive this condition is on a hyperbolic paraboloid of one sheet in E^3 is still an open question.

Chapter 5

Theory of Spherical Biarc Interpolation

The curve interpolation problem on the unit sphere $S^3 \subset E^4$ arises in computer animation for controlling the orientation of objects. This problem motivates us to consider the following interpolation problem on a sphere in E^d, $d \geq 3$. Given two points and their associated tangent directions on a sphere, we investigate the existence of *spherical biarcs* to interpolate the given points and the tangent directions on the sphere. *Spherical biarcs* are curves consisting of two smoothly joining circular arcs on a sphere. Therefore they can be easily represented and evaluated. It is shown that there always exist spherical biarcs interpolating any data as specified above, and a complete description of all such spherical biarcs is given. Specifically, for generic data on a sphere S in E^d it is shown that there is a family of interpolating spherical biarcs with one degree of freedom; for the class of singular data on S, the family of the interpolating spherical biarcs has $d - 2$ degrees of freedom. Several more properties of spherical biarcs are derived in this chapter.

The problem dealt with in this chapter is a special case of Section 4.3, where the quadric is a sphere; therefore there are similarities between the discussion here and that of Section 4.3. The most important distinction between the spherical biarcs to be discussed here and those investigated in Section 4.3 is that the restriction that a biarc is composed of two rational quadratic Bézier curves with positive weights no longer applies here; that is, here we will discuss spherical biarcs in the most general setting.

5.1 Introduction

A biarc is a curve composed of two circular arcs joining smoothly with unit tangent vector continuity. Biarcs in the plane and in 3-D space have been studied for curve and surface modeling in CAGD [Bez72, Bol75, SuL89, Sab76, Sha87, NuM88]. The main reasons for using biarcs are that it is easy to represent a circular arc and it is simple to compute points on it. Also, on an NC milling machine the circular arc is usually the only curved locus that can be traced efficiently [Sab76]. In this chapter we will study the biarc on a sphere in E^d, $d \geq 3$, which will be called the *spherical biarc*. To distinguish it from biarcs in the plane or in space, the latter will be called *plane* or *space biarcs*.

The goal of this chapter is to study interpolation properties of spherical biarcs. An application of the interpolation of a point sequence on a sphere arises in computer animation, when a smooth motion needs to be interpolated from the positions of the object at a series of distinct time instants. This problem, ignoring the translation of the object positions, can be transformed to an interpolation problem for unit quaternions, which form the unit sphere $S^3 \subset E^4$ under Euclidean topology. It is well known [Sho85] that any rotation in E^3 can be represented by two diametrically opposite unit quaternions on $S^3 \subset E^4$, and conversely, a unit quaternion determines a unique rotation in E^3. Any two orientations of an object can be made to coincide through a rotation in E^3. So orientations can be represented by unit quaternions with respect to a reference orientation. Thus the orientation part of the motion interpolation problem is reduced to an interpolation problem on $S^3 \subset E^4$. Note that the interpolation of translation in the motion interpolation problem can be solved relatively easily, e.g. using a cubic spline. For more details about this application the reader is referred to [Sho85]. See also the next chapter.

Several interpolation schemes have been proposed in the literature to solve the above spherical interpolation problem. In [Sho85] Shoemake constructs a spherical interpolating curve in a way analogous to the deCasteljau construction for the cubic Bézier curve. In [Sho87] Shoemake proposes another spherical interpolant whose construction is analogous to Boehm's quadrilateral construction for the cubic curve. Only exponential parametric representations are known for curves generated by these two constructions. In [GeR91] the following method is proposed: First cubic polynomial curves are used to interpolate spherical data points. Since the resulting curves do not in general lie on the sphere, a central projection from the center of the sphere is applied to project the curve onto the sphere. While the parametric representation of such curves is relatively simple, the shape properties of these curves are yet to be studied. None of the above spherical interpolants are rational curves.

In this chapter we consider the following spherical biarc interpolation problem: Let a data set be $D = \{X_0, T_0, X_1, T_1\}$, where X_0, X_1 are two distinct points on a sphere S in E^d, $d \geq 3$, and T_0, T_1 are two tangent directions defined at X_0 and X_1, respectively. Find a spherical biarc B on S to interpolate the points X_0, X_1, and the tangent directions T_0 and T_1. Such a biarc B is said to interpolate the data D. Our main conclusion is that there always exist spherical biarcs to interpolate any such data D. So we will always have an interpolating rational spline curve since circular arcs are rational quadratic curves.

The main results and the basic organization of this chapter are as follows. Let S be a sphere in E^d, $d \geq 3$. It will be shown that for generic data D on S, to be called regular data, there exist a family of spherical biarcs interpolating D with one degree of freedom. For a class of special data on S, to be called singular data, there exist a family of spherical biarc interpolants with $d-2$ degrees of freedom. These results will be derived in Section 5.3 after preliminaries are reviewed in Section 5.2.

In Section 5.4 more properties of spherical biarcs are derived. The most important one is that the locus of all joints of spherical biarcs interpolating regular data D is a circle on S^{d-1}, while the locus for singular data is a $d-2$ dimensional sphere on S^{d-1}. Section 5.5 contains concluding remarks.

5.2 Preliminaries

Using homogeneous coordinates, a real sphere in E^d can be represented by the quadratic equation $X^T A X = 0$, with

$$A = \begin{bmatrix} I_d & b \\ b^T & \alpha \end{bmatrix},$$

where I_d is the $d \times d$ identity matrix, and $b^T b - \alpha = r^2 > 0$, where $r > 0$ is the radius of the sphere. Since the properties we shall deal with on a sphere in E^d can always be transformed affinely onto the unit sphere $S^{d-1} \subset E^d$ without affecting the nature of the problems and solutions, we shall always assume that the sphere under discussion is the unit sphere $S^{d-1} \subset E^d$ with equation $X^T A X = 0$, where

$$A = \begin{bmatrix} I_d & 0 \\ 0 & -1 \end{bmatrix}.$$

One consequence of this assumption is that $X^T A Y = X^T Y$ if X or Y is a point at infinity. This fact will be used frequently later on.

The tangent hyperplane of S^{d-1} at point $X_0 \in S^{d-1}$ is the polar hyperplane of X_0 with respect to S^{d-1}: $X_0^T AX = 0$. Also we say that a point $X_0 \in E^d$ is *inside* S^{d-1} if $X_0^T AX_0 < 0$ and *outside* S^{d-1} if $X_0^T AX_0 > 0$.

In the following discussion all circular arcs will be represented as rational quadratic Bézier curves. It is well known that any circular arc can be represented in the *standard rational quadratic Bézier form* [Lee87, Pat86], which has the homogeneous representation

$$P(u) = P_0 B_{0,2}(u) + wP_1 B_{1,2}(u) + P_2 B_{2,2}(u), \quad 0 \le u \le 1, \tag{5.1}$$

where P_0,P_1,P_2 are the *control points* of the curve, and P_0 and P_2 are assumed to be in normalized form. When P_1 is a finite point, it is assumed to be in normalized form. We will also refer to $P_0P_1P_2$ as the *control polygon* of the curve. The scalar w is called the *weight* of the curve, and $B_{i,2}(u) = \frac{2}{i!(2-i)!}u^i(1-u)^{2-i}$, $i = 0,1,2$, are the *second degree Bernstein polynomials.*

By the symmetry of a circular arc, it is easy to see that a necessary condition for (5.1) to represent a circular arc is that $P_0P_1P_2$ is an isosceles triangle with the base $\overline{P_0P_2}$ when P_1 is a finite point or the direction represented by P_1 is perpendicular to $\overline{P_0P_2}$ when P_1 is a point at infinity. Besides, the weight must be chosen appropriately. This is done as follows when the arc is restricted to lie on the sphere S^{d-1}. Let C be a circular arc on S^{d-1} given by (5.1). Then $P(u)^T AP(u) = 0$. Since it is required that P_0, P_2 be on S^{d-1} and the straight lines P_0P_1 and P_1P_2 be tangent to S^{d-1} at P_0 and P_2, respectively, $P_0^T AP_0 = P_2^T AP_2 = P_0^T AP_1 = P_2^T AP_1 = 0$. Therefore substituting (5.1) in $P(u)^T AP(u) = 0$, we have

$$2P_0^T AP_2 B_{0,2}(u)B_{2,2}(u) + w^2 P_1^T AP_1 B_{1,2}^2(u) = 0.$$

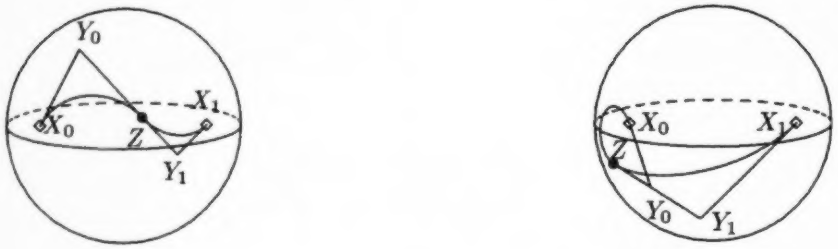
Since $B_{1,2}^2(u) = 4B_{0,2}(u)B_{2,2}(u)$ and $P_1^T AP_1 \ne 0$, we obtain

$$w^2 = -\frac{P_0^T AP_2}{2P_1^T AP_1}. \tag{5.2}$$

As P_0 and P_2 are in normalized form,

$$P_0^T AP_2 = -\frac{1}{2}(P_0 - P_2)^T A(P_0 - P_2) < 0.$$

On the other hand, $P_1^T AP_1 > 0$ since P_1 is lies outside the sphere. So for a well posed problem, two real nonzero values of w with opposite signs can be solved for from (5.2), which will be referred to as the *positive* weight and the *negative* weight of the curve. These two weights give respectively the minor arc and the major arc of the circle on S^{d-1} with control polygon $P_0P_1P_2$. A minor arc lies inside the triangle $\Delta P_0P_1P_2$, while the complementary major arc lies outside $\Delta P_0P_1P_2$.



(*a*) (*b*)

Figure 5.2.1 The two cases indicated in Lemma 5.2.1.

Semicircles are represented by (5.1) when P_1 is a point at infinity. In this case the positive weight and the negative weight give two mutually complementary semicircles. Apparently, a weight w solved for from (5.2) can never be zero unless $[P_0] = [P_2]$.

The condition on the smooth joining of two circular arcs in Bézier form (5.1) is important in constructing a spherical biarc. Let C_0 and C_1 be two circular arcs on S^{d-1} joining at Z with control polygons X_0Y_0Z and ZY_1X_1, respectively. Then a necessary condition for them to meet with unit tangent vector continuity is that Y_0, Y_1 and Z are collinear. If we assume that this is the case, the following conditions for their smooth joining, which are given without proof, are obvious.

Lemma 5.2.1: *Using the above notation, assume that* Y_0, Y_1 *and* Z *are collinear,* $[Z] \neq [Y_0]$ *and* $[Z] \neq [Y_1]$. *When* Z *is interior to the line segment* $\overline{Y_0Y_1}$, *assuming* $[Y_0] \neq [Y_1]$, C_0 *and* C_1 *join smoothly at* Z *if and only if they both simultaneously have positive weights or negative weights; when* Z *is outside* $\overline{Y_0Y_1}$, *including the case* $[Y_0] = [Y_1]$, C_0 *and* C_1 *join smoothly if and only if one of* C_0 *and* C_1 *has a positive weight and the other has a negative weight.*

The two cases indicated in Lemma 5.2.1 are illustrated in Figure 5.2.1. Note that Lemma 5.2.1 is also true when Y_0 or Y_1 are points at points at infinity. When only one of them is a point at infinity, the segment $\overline{Y_0Y_1}$ becomes a half-line; when both of Y_0 and Y_1 are points at infinity, which are necessarily two opposite directions since Z is a finite point, $\overline{Y_0Y_1}$ becomes the straight line passing through Z and Y_0.

When C_0 and C_1 form a biarc B, the points X_0, Y_0, Y_1 and X_1 are called the *control points* of B. And Z is called the *joint* of B.

5.3 Existence of spherical biarc interpolants

In this section we will prove the existence of spherical biarcs interpolating any two distinct points X_0, X_1 and the associated tangent directions T_0, T_1 on the sphere $S^{d-1} \subset E^d$, $d \geq 3$. We emphasize again that the restriction to positive weights of a biarc used in the previous chapter no longer applies here. In order to define a tangent direction at a point of a biarc so as to make it meaningful to say that a tangent direction is interpolated, we associate a biarc passing through X_0 and X_1 with the direction in which a variable point on the biarc moves from X_0 to X_1 through the joint point. For our purpose we will distinguish two kinds of data $D = \{X_0, X_1, T_0, T_1\}$ to be interpolated; the first is called regular data, which is generic, and the second kind is called singular data (definitions will be given later). In Subsection 5.3.1 we treat the case of regular data and show that there exists a family of biarcs interpolating any regular data, and this family has one degree of freedom. In Subsection 5.3.2 we prove the existence of spherical biarcs for singular data and show that there exists a family of spherical biarcs interpolating any singular data, and this family has $d-2$ degrees of freedom.

5.3.1 Existence for regular data

Let X_0, X_1 be two distinct points in normalized form on $S^{d-1} \subset E^d$. Let T_0, T_1 be two tangent directions specified at X_0, X_1 respectively. To have a well defined problem, T_0 and T_1 are required to be tangent to S^{d-1} at X_0 and X_1, respectively. Thus $X_0^T A X_0 = X_0^T A T_0 = X_1^T A X_1 = X_1^T A T_1 = 0$. Furthermore, without loss of generality, assume that $T_0^T A T_0 = T_1^T A T_1 = 1$. Any data $D = \{X_0, T_0, X_1, T_1\}$ on S^{d-1} satisfying the above assumptions is called a *data set* for the biarc problem, or *data* for short.

Our goal now is to find a biarc B on S^{d-1} to interpolate the data D. B is also said to be a biarc for D. Let biarc B be composed of two arcs C_0 and C_1 with control polygons X_0Y_0Z and ZY_1X_1, where Z is the joint. Let

$$\begin{aligned} Y_0 &= X_0 + k_0 T_0, \\ Y_1 &= X_1 - k_1 T_1, \end{aligned} \tag{5.3}$$

where k_0, k_1 are parameters to be determined.

Definition 5.3.1: *A biarc is* degenerate *if one of its arcs reduces to a single point. A biarc that is not degenerate is called* proper.

By definition, the spherical biarc B is degenerate if and only if $[Z] = [X_0]$ or $[Z] = [X_1]$.

Necessity follows from the definition. To see the sufficiency, without loss of generality, suppose that $[Z] = [X_0]$. In this case we obtain $w = 0$ from (5.2). Therefore the Bézier curve becomes the line segment $\overline{ZX_0}$, which collapses to a point. Only proper spherical biarcs interpolating D are useful to us because in a degenerate biarc the tangent at the end point of the degenerate arc is not defined. So we will assume $k_0k_1 \neq 0$ for otherwise from (5.3) $[Y_0] = [X_0]$ or $[Y_1] = [X_1]$, consequently, $[Z] = [X_0]$ or $[Z] = [X_1]$, and the resulting biarc would be degenerate since the control polygon reduces to a point. Another observation concerning the signs of k_0 and k_1 is that, in order to interpolate the end tangent directions T_0 and T_1, the arc C_i of B, $i = 0, 1$, must take the positive weight if $k_i > 0$ or the negative weight if $k_i < 0$. Hence C_i is a minor arc if and only if $k_i > 0$ and Y_i is a finite point.

We presently assume that the points Y_0 and Y_1 are distinct; this assumption will be justified later. Thus the straight line Y_0Y_1 is uniquely defined. Since the point Z is on Y_0Y_1, let $Z = \lambda Y_0 + \mu Y_1$. Then Y_0Y_1 must be tangent to S^{d-1} at Z. So the Joachimsthal's equation [SeK52], obtained by substituting $Z = \lambda Y_0 + \mu Y_1$ in sphere equation $X^TAX = 0$,

$$\lambda^2 Y_0^T A Y_0 + 2\lambda\mu Y_0^T A Y_1 + \mu^2 Y_1^T A Y_1 = 0$$

has double roots λ/μ. Thus the discriminant

$$\Delta \equiv 4[(Y_0^T A Y_1)^2 - (Y_0^T A Y_0)(Y_1^T A Y_1)] = 0. \tag{5.4}$$

From (5.3), since $X_0^T A X_0 = X_0^T A T_0 = 0$ and $T_0^T A T_0 = 1$,

$$Y_0^T A Y_0 = X_0^T A X_0 + 2k_0 X_0^T A T_0 + k_0^2 T_0^T A T_0 = k_0^2.$$

Similarly, $Y_1^T A Y_1 = k_1^2$. Thus by (5.4), $(Y_0^T A Y_1)^2 - k_0^2 k_1^2 = 0$. That is,

$$Y_0^T A Y_1 = k_0 k_1 \text{ or} \tag{5.5}$$

$$Y_0^T A Y_1 = -k_0 k_1. \tag{5.6}$$

On the other hand, when $\Delta = 0$, the double root of the Joachimsthal's equation is

$$\lambda/\mu = -(Y_0^T A Y_1)/(Y_0^T A Y_0).$$

Thus, omitting a nonzero multiplicative factor, the tangent point Z of the line Y_0Y_1 with S^{d-1} is

$$Z = (Y_0^T A Y_1) Y_0 - (Y_0^T A Y_0) Y_1. \tag{5.7}$$

Substituting (5.5) or (5.6) in (5.7), respectively, we have

$$Z = k_0 k_1 Y_0 - k_0^2 Y_1 \text{ or}$$

$$Z = -k_0 k_1 Y_0 - k_0^2 Y_1,$$

which can be rewritten as

$$Z = k_1 Y_0 - k_0 Y_1 \text{ or} \tag{5.8}$$

$$Z = k_1 Y_0 + k_0 Y_1, \tag{5.9}$$

since $k_0 \neq 0$. We note that (5.8) and (5.9) are derived from (5.5) and (5.6), respectively.

So far we have proved the following lemma.

Lemma 5.3.1: *For any data $D = \{X_0, T_0, X_1, T_1\}$, if a proper spherical biarc interpolating D exists with the control points Y_0 and Y_1 being distinct, then the joint Z is given by (5.8) or (5.9) for some k_0 and k_1 satisfying (5.5) or (5.6), respectively.*

For simplicity, we adopt the following convention: when we say that the solutions k_0 and k_1 of Eqn. (5.6) (or Eqn. (5.5)) give a biarc, we mean that the control points Y_0 and Y_1 are given by (5.3) and the joint Z is determined by (5.9) (or (5.8)).

Now we need to decide which of (5.8) and (5.9) gives a smooth joint Z. To this end we introduce the definition of singular data.

Definition 5.3.2: *The data $D = \{X_0, T_0, X_1, T_1\}$ is* singular *if*

$$X_0 + \rho T_0 = X_1 + \rho T_1$$

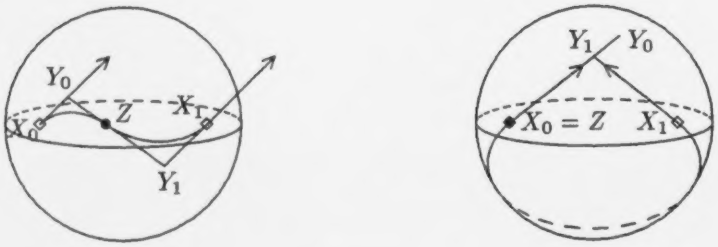
for some $\rho \neq 0$ or if $T_0 = T_1$. The data that is not singular is called regular.

Fig. 5.3.1 illustrates instances of regular data and singular data.

A necessary, but not sufficient, condition for D to be singular is that X_0, X_1, T_0 and T_1 are coplanar, i.e. they are linearly dependent. A simple classification of coplanar data D can be obtained easily, which will be used in later discussions. For coplanar D, considering the last components of X_0, T_0, X_1, and T_1, we have either (1) $T_0 = \rho T_1$ for some $\rho \neq 0$; or (2) $X_0 + \rho_0 T_0 = X_1 + \rho_1 T_1$ for $\rho_0 \neq 0$, $\rho_1 \neq 0$. In case (1), by $T_0^T A T_0 = T_1^T A T_1 = 1$, we have either $T_0 = T_1$, which stands for singular data, or $T_0 = -T_1$, which stands for regular data. In case (2), from

$$(X_0 + \rho_0 T_0)^T A (X_0 + \rho_0 T_0) = (X_1 + \rho_1 T_1)^T A (X_1 + \rho_1 T_1),$$

and using $X_0^T A X_0 = X_0^T A T_0 = X_1^T A X_1 = X_1^T A T_1 = 0$, $T_0^T A T_0 = T_1^T A T_1 = 1$, we obtain $\rho_0^2 = \rho_1^2$, i.e. $\rho_0 = \rho_1$ or $\rho_0 = -\rho_1$. So we have either $X_0 + \rho_0 T_0 = X_1 + \rho_0 T_1$, which stands for singular data, or $X_0 + \rho_0 T_0 = X_1 - \rho_0 T_1$, which stands for regular data.



(a) (b)

Figure 5.3.1 Regular data and singular data: (a) Regular data on S^2 with a proper biarc; (b) Singular data on S^2 with a degenerate biarc. Note that the two tangents in (b) intersect at a point in E^3.

In this subsection we will focus on the existence of biarcs interpolating regular data. When the data D is regular, the next lemma justifies the earlier assumption that $[Y_0] \neq [Y_1]$ when Y_0, Y_1 are control points of a proper biarc for D.

Lemma 5.3.2: If X_0Y_0Z and ZY_1X_1 are the control polygons of a proper biarc B interpolating regular data D, then $[Y_0] \neq [Y_1]$.

PROOF: Suppose there is a proper biarc interpolating D with $[Y_0] = [Y_1]$. First assume that $[Y_0] = [Y_1]$ is a finite point. In this case D is coplanar data. By the above classification of coplanar data and the fact that D is regular, we have

$$Y_0 = X_0 + \rho_0 T_0 = X_1 - \rho_0 T_1 = Y_1,$$

for some $\rho_0 \neq 0$. Since $k_0 = k_1 = \rho_0 \neq 0$, in order to interpolate T_0 and T_1 at X_0 and X_1, C_0 and C_1 must take positive weights or negative weights simultaneously, depending on the sign of ρ_0. On the other hand, the joint Z must be the tangent point to S^{d-1} of a straight line passing through Y_0, which is outside S^{d-1}. Thus Z is outside the degenerate line segment $\overline{Y_0Y_1}$ on the line Y_0Z. By Lemma 5.2.1, C_0 and C_1 cannot join smoothly. This is a contradiction.

Now consider the case where $[Y_0] = [Y_1]$ is a point at infinity. Again from the classification of coplanar data, we have $T_0 = -T_1$ since D is regular. In this case both C_0 and C_1 are semicircles. Since C_0, C_1 interpolate tangent directions T_0 and T_1 at X_0 and X_1, respectively, the tangent directions of C_0 and C_1 at joint Z are $-T_0$ and $-T_1 = T_0$, respectively. Therefore C_0 and C_1 cannot join smoothly at Z since they have opposite tangent directions at Z. □

By Lemma 5.3.1 and Lemma 5.3.2, we have

Lemma 5.3.3: *For regular data D, the joint Z of any proper spherical biarc interpolating D is given by (5.8) or (5.9).*

Now we assert that a point Z is the joint of some proper biarc interpolating regular data D if it is given by (5.9), where k_0 and k_1, with $k_0k_1 \neq 0$, satisfy (5.6), i.e.

$$(X_0^T A X_1) + (X_1^T A T_0)k_0 - (X_0^T A T_1)k_1 + (1 - T_0^T A T_1)k_0k_1 = 0. \tag{5.10}$$

Theorem 5.3.4: *For regular data $D = \{X_0, T_0, X_1, T_1\}$, a point Z given by (5.9) with any solutions k_0, k_1 of (5.10), where $k_0k_1 \neq 0$, is the joint of a proper spherical biarc interpolating D, and the associated control points Y_0 and Y_1 are determined by k_0, k_1 through (5.3).*

We need the following lemmas to prove this theorem.

Lemma 5.3.5: *Given data $D = \{X_0, T_0, X_1, T_1\}$ on S^{d-1}, $[Y_0] = [Y_1]$ for some k_0 and k_1 satisfying (5.10) if and only if D is singular, where Y_0 and Y_1 are given by (5.3).*

The proof for Lemma 5.3.5 is essentially the same as that for Lemma 4.3.2 of Chapter 4, so it is omitted here.

Lemma 5.3.6: *Let D be regular data on S^{d-1}. Let k_0 and k_1 be any solutions of (5.10) with $k_0k_1 \neq 0$. Then $[Z] \neq [X_0]$ and $[Z] \neq [X_1]$, where Z is given by (5.9).*

PROOF: The proof is by contradiction. Suppose the opposite. Without loss of generality, assume that for regular data $D = \{X_0, T_0, X_1, T_1\}$ there exist solutions k_0 and k_1 of (5.10) with $k_0k_1 \neq 0$ such that $Z = \beta X_0$ for some $\beta \neq 0$. Then by (5.9) and (5.3),

$$\beta X_0 = k_1(X_0 + k_0T_0) + k_0(X_1 - k_1T_1).$$

Since X_0 and X_1 are in normalized form and the last components of T_0 and T_1 are zero, $\beta = k_0 + k_1$. Thus

$$k_0X_0 - k_0k_1T_0 = k_0X_1 - k_0k_1T_1,$$

or, since $k_0 \neq 0$,

$$X_0 - k_1T_0 = X_1 - k_1T_1.$$

Since $k_1 \neq 0$, by definition, D is singular. This is a contradiction. □

PROOF of Theorem 5.3.4: By Lemma 5.3.5, given regular data D, $[Y_0] \neq [Y_1]$ for any k_0 and k_1 satisfying (5.10). By Lemma 5.3.6, $[Z] \neq [X_0]$ and $[Z] \neq [X_1]$ for any solutions k_0 and k_1 of (5.10) with $k_0 k_1 \neq 0$; therefore no degenerate biarc can arise from solutions of (5.10) when $k_0 k_1 \neq 0$. Now we will show by construction that a point Z given by (5.9), with k_0 and k_1 satisfying (5.10) and $k_0 k_1 \neq 0$, is the joint of a proper biarc on S^{d-1} interpolating D. Let C_0 and C_1 be the two arcs with control polygons $X_0 Y_0 Z$ and $Z Y_1 X_1$, respectively. There are four cases to consider.

(*i*) $k_0 > 0$ and $k_1 > 0$. Since $[Y_0] \neq [Y_1]$ and Y_0 and Y_1 are in normalized form, by (5.9), Z is between Y_0 and Y_1. To interpolate T_0 and T_1 at the two end points, let C_0 and C_1 both take positive weights. Then by Lemma 5.2.1, we have a smooth biarc on S^{d-1} interpolating D. This case is illustrated in Fig. 5.3.2(*a*).

(*ii*) $k_0 < 0$, $k_1 < 0$. As in the first case, $[Y_0] \neq [Y_1]$ and Z is between Y_0 and Y_1. To interpolate T_0 and T_1 at the two end points, we assign the negative weights to C_0 and C_1. By Lemma 5.2.1, the two arcs join smoothly at Z. This case is illustrated in Fig. 5.3.2(*b*).

(*iii*) $k_0 > 0$ and $k_1 < 0$. By (5.9), Z is outside the segment $\overline{Y_0 Y_1}$. There are now two subcases: when $k_0 + k_1 > 0$, Y_1 is between Z and Y_0; when $k_0 + k_1 < 0$, Y_0 is between Z and Y_1. In both cases, to interpolate T_0 and T_1, let C_0 take the positive weight and C_1 take the negative weight. Then by Lemma 5.2.1, we have a smooth biarc interpolating D. This case is illustrated in Fig. 5.3.2(*c*) and (*d*). We remark that $k_0 + k_1 \neq 0$ always holds under the assumption of the theorem, because $k_0 + k_1$ is the last component of the finite point $Z \in S^{d-1}$. This fact will also be proved later in Theorem 5.4.4.

(*iv*) $k_0 < 0$ and $k_1 > 0$. In this case, by (5.9), Z is outside the segment $\overline{Y_0 Y_1}$. There are again two subcases: when $k_0 + k_1 > 0$, Y_0 is between Z and Y_1; when $k_0 + k_1 < 0$, Y_1 is between Z and Y_0. In both cases, to interpolate T_0 and T_1, let C_0 take the negative weight and C_1 take the positive weight. Then by Lemma 5.2.1, we have a smooth biarc interpolating D. This case is illustrated in Fig. 5.3.2(*e*) and (*f*). As remarked in case (*iii*), $k_0 + k_1 \neq 0$ always holds for regular data.

In the above cases it is implicitly assumed that k_0 and k_1 are finite. For regular data, the solutions k_0 and k_1 of Eqn. (5.10) cannot both be infinite. For from (5.10) we have

$$\frac{X_0^T A X_1}{k_0 k_1} + \frac{X_1^T A T_0}{k_1} - \frac{X_0^T A T_1}{k_0} + 1 - T_0^T A T_1 = 0.$$

If both k_0 and k_1 are infinite then it would follow that $1 - T_0^T A T_1 = 0$; thus $T_0 = T_1$, contradicting the assumption that D is regular. Now suppose only one of k_0 and k_1 is infinite. Then simply replacing the segment $\overline{Y_0 Y_1}$ by a half-line $\overrightarrow{Y_0 Y_1}$ or $\overrightarrow{Y_1 Y_0}$ in the above cases furnishes the desired proof. Hence in all cases, under the assumptions of the theorem, a proper biarc can be constructed to interpolate regular data D. □

By Theorem 5.3.4, the existence of proper biarcs interpolating D depends on the existence of nonzero solutions k_0 and k_1 of Eqn. (5.10). The next lemma shows that for regular data Eqn. (5.10) always has such solutions. In the following ∞^n denotes the number of elements in a family with n degrees of freedom, where n is a nonnegative integer.

Lemma 5.3.7: *For regular data $D = \{X_0, T_0, X_1, T_1\}$, Eqn. (5.10) has ∞^1 many pairs of nonzero solutions k_0, k_1, i.e. $k_0 k_1 \neq 0$.*

PROOF: All solution pairs k_0, k_1 of (5.6) comprise points on the curve represented by (5.10) in the k_0-k_1 plane. We just have to show that there are ∞^1 points of the curve which are not on the k_0 or k_1 axis.

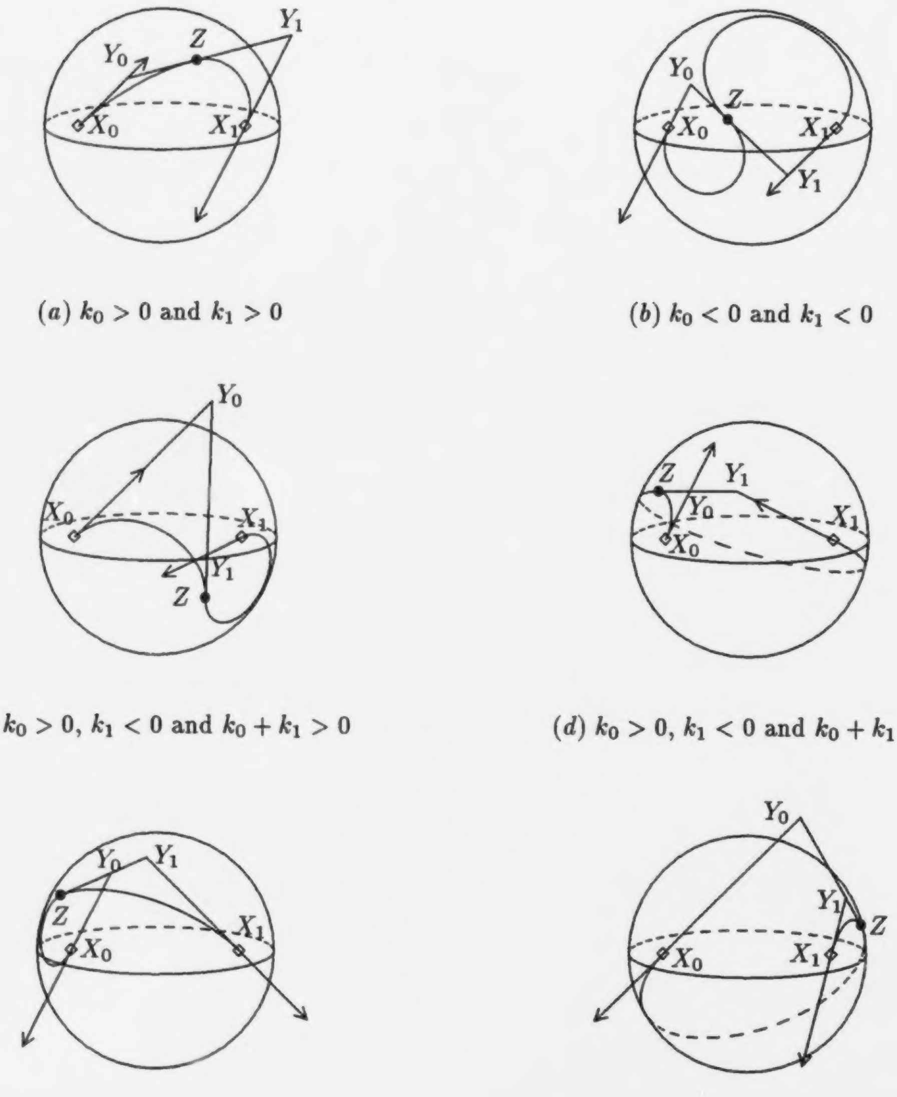
Since X_0 and X_1 are distinct, $X_0^T A X_1 \neq 0$. Also we have $1 - T_0^T A T_1 = \frac{1}{2}(T_0 - T_1)^T A (T_0 - T_1) > 0$, since $T_0 \neq T_1$ as D is regular. So the constant term and the coefficient of $k_0 k_1$ in Eqn. (5.10) are nonzero. This implies that Eqn. (5.10) represents hyperbola and $(k_0, k_1) = (0, 0)$ is not a solution of Eqn. (5.10).

If Eqn. (5.10) is irreducible, it represents a proper hyperbola in the $k_0 k_1$ plane. Therefore there are ∞^1 many points (k_0, k_1) of the hyperbola for which $k_0 k_1 \neq 0$. If Eqn. (5.10) is reducible, it represents two straight lines, at least one of which is not the k_0 or k_1 axis; for otherwise the origin $(k_0, k_1) = (0, 0)$ would be a point on the curve represented by (5.10), a contradiction. So again there are ∞^1 points (k_0, k_1) on the two lines for which $k_0 k_1 \neq 0$. In fact, we will later show that Eqn. (5.10) is irreducible for any regular data. □

Theorem 5.3.4 indicates the following way to construct a smooth biarc with a joint given by (5.9). For each of the arcs C_0 and C_1 there are two possible weights that can be computed from (5.2); the positive weight is used if $k_i > 0$ and the negative weight is used if $k_i < 0$, $i = 0, 1$.

Combining Theorem 5.3.4 and Lemma 5.3.7, we have proved the existence of proper spherical biarcs interpolating regular data D. But for a complete description of all proper biarcs for a given regular data set D, by Lemma 5.3.1, we still have to consider whether Eqn. (5.8) gives any point Z that is the joint of a proper biarc interpolating D when $[Y_0] \neq [Y_1]$. Recall that the case where $[Y_0] = [Y_1]$ has been excluded by Lemma 5.3.2.

Lemma 5.3.8: *For any data D, a point Z given by (5.8) with k_0 and k_1 satisfying (5.5) is not the joint of any proper spherical biarc interpolating D with $[Y_0] \neq [Y_1]$.*


Y_1
Z
Y_0
X_0
X_1
Y_0
Z
X_0
X_1
Y_1
Y_0
X_0
X_1
Y_1
Z
Z
Y_1
Y_0
X_1
X_0
Y_0
Y_1
Z
X_1
X_0
Y_0
Y_1
Z
X_0
X_1

(*a*) $k_0 > 0$ and $k_1 > 0$

(*b*) $k_0 < 0$ and $k_1 < 0$

(*c*) $k_0 > 0$, $k_1 < 0$ and $k_0 + k_1 > 0$

(*d*) $k_0 > 0$, $k_1 < 0$ and $k_0 + k_1 < 0$

(*e*) $k_0 < 0$, $k_1 > 0$ and $k_0 + k_1 > 0$

(*f*) $k_0 < 0$, $k_1 > 0$ and $k_0 + k_1 < 0$

Figure 5.3.2 The six cases in the proof of Theorem 5.3.4.

PROOF: The proof is by contradiction. Let k_0, k_1 be solutions of Eqn. (5.5) and $k_0k_1 \neq 0$, recalling that $k_i = 0$, $i = 0$ or 1, leads to a degenerate biarc. Let C_0 and C_1 be the arcs with control polygons X_0Y_0Z and ZY_1X_1, respectively, where Y_0 and Y_1 are given by (5.3) and Z by (5.8), with the above k_0 and k_1. Suppose that C_0 and C_1 form a proper biarc B interpolating D.

$[Y_0] \neq [Y_1]$ implies that the straight line Y_0Y_1 is uniquely defined, and the argument leading to (5.8) is valid, i.e. any possible joint must be given by (5.8). Assume that $[Z] \neq [X_0]$ and $[Z] \neq [X_1]$, for otherwise a degenerate biarc would result. Now there are four cases to consider: (i) $k_0 > 0$ and $k_1 > 0$; (ii) $k_0 < 0$ and $k_1 < 0$; (iii) $k_0 > 0$ and $k_1 < 0$; (iv) $k_0 < 0$ and $k_1 > 0$. We will consider just case (i). The proofs for the remaining cases are similar and thus omitted.

(i) $k_0 > 0$ and $k_1 > 0$. First assume that k_0 and k_1 are finite. In order for B to interpolate T_0 and T_1 at X_0 and X_1, respectively, we are forced to assign the positive weights to both arcs C_0 and C_1. On the other hand, by (5.8), Z is outside the segment $\overline{Y_0Y_1}$ when $k_0 > 0$ and $k_1 > 0$. Thus by Lemma 5.2.1, C_0 and C_1 do not meet smoothly at Z. This is a contradiction. When only one of k_0 and k_1 is infinite, simply replace the segment $\overline{Y_0Y_1}$ by a ray, and a contradiction still follows from Lemma 5.2.1. When both k_0 and k_1 are infinite, $[Y_0] = [Y_1]$, which is the case excluded by the lemma. □

The following is the complete existence theorem for proper spherical biarcs interpolating regular data.

Theorem 5.3.9: *Let D be regular data on $S^{d-1} \subset E^d$, $d \geq 3$. Then there are ∞^1 many proper spherical biarcs interpolating D on S^{d-1}. The joints of all these biarcs are given by (5.9) with k_0 and k_1 satisfying (5.10) and $k_0k_1 \neq 0$. The control points Y_0 and Y_1 are given by (5.3) in terms of k_0 and k_1.*

In Section 5.4 it will be shown that for regular data D all the joints of proper biarcs for D together with X_0 and X_1 form a circle on S^{d-1}. Fig. 5.3.3 shows a family of proper biarcs interpolating the indicated regular data.

5.3.2 Existence for singular data

In this subsection we consider the existence of proper biarcs interpolating singular data D on S^{d-1}.

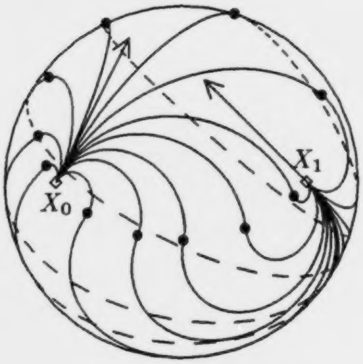


Figure 5.3.3 A family of biarcs on S^2 interpolating regular data.

Theorem 5.3.10: *Let D be singular data on S^{d-1}. Let $Y = T_0$ if $T_0 = T_1$ or $Y = X_0 + \rho T_0$ if $X_0 + \rho T_0 = X_1 + \rho T_1$ for some $\rho \neq 0$. Let J be the $(d-2)$-dimensional sphere which is the intersection of S^{d-1} with the polar hyperplane of Y with respect to S^{d-1}, i.e. $J = \{X | X^T AX = 0,\ Y^T AX = 0\}$. Then any point $Z \in J - \{X_0, X_1\}$ is the joint of a proper spherical biarc interpolating D with control points $Y_0 = Y_1 = Y$.*

PROOF: When $[Y_0] = [Y_1] = [Y]$, in order to construct a biarc consisting of two circular arcs C_0 and C_1 with control polygons X_0Y_0Z and ZY_1X_1 respectively, we need to find a joint Z which is the tangent point to S^{d-1} of a straight line passing through Y_0. All such points Z form the $(d-2)$-dimensional sphere J, which is defined by $X^T AX = 0$ and $Y_0^T AX = 0$.

It is easily verified that X_0 and $X_1 \in J$. They cannot be the joint of a proper biarc interpolating D. For any $Z \in J - \{X_0, X_1\}$, we need to show that there is a proper spherical biarc for D with joint Z. Note that each such Z is outside the degenerate line segment $\overline{Y_0Y_1}$ on the line ZY_0. First suppose that $Y = Y_0 = Y_1$ is a finite point. In this case $k_0 = \rho$, $k_1 = -\rho$. So, to meet the end tangent conditions, we must assign the positive weight to C_0 and the negative weight to C_1 if $\rho > 0$, and the opposite signs if $\rho < 0$. Then by Lemma 5.2.1, C_0 and C_1 join smoothly with such assignments of weights.

Now suppose that $[Y] = [Y_0] = [Y_1]$ is a point at infinity, in which case $T_0 = T_1$ and C_0 and C_1 are semicircles. We choose the weights of C_0 and C_1 so that T_0 and T_1 are interpolated. Then C_0 and C_1 have $-T_0$ and $-T_1 = -T_0$ as tangent directions, respectively, at the joint Z. So C_0 and C_1 join smoothly at Z. □

Theorem 5.3.10 only covers the case where $[Y_0] = [Y_1]$. In order to give a complete description of all proper biarcs interpolating singular data D, by Lemma 5.3.1, we have to consider whether (5.8) or (5.9) can yield a joint of some proper biarc interpolating singular data D when $[Y_0] \neq [Y_1]$. Lemma 5.3.8 indicates that (5.8) does not give the joint of a proper biarc for any data, whether it is regular or singular. So in the case where $[Y_0] \neq [Y_1]$ and the data D is singular, we only need to consider (5.9) with k_0 and k_1 satisfying (5.6).

The next lemma states that for singular data D only degenerate biarcs result from the solution k_0, k_1 of (5.6). Therefore all proper biarcs for singular data are completely described by Theorem 5.3.10.

Lemma 5.3.11: *For singular data D on S^{d-1}, any solution k_0 and k_1 of (5.10) for which $[Y_0] \neq [Y_1]$ yields only degenerate biarcs.*

PROOF: Let $D = \{X_0, T_0, X_1, T_1\}$ be singular. There are two cases to consider: (i) $X_0 + \rho T_0 = X_1 + \rho T_1$ for some finite $\rho \neq 0$; and (ii) $T_0 = T_1$.

(i) $X_0 + \rho T_0 = X_1 + \rho T_1$ for some $\rho \neq 0$. In this case $T_0 \neq T_1$. Then the left hand side of Eqn. (5.10) becomes

$$\begin{aligned} & X_0^T A(X_0 + \rho T_0 - \rho T_1) + k_0(X_0 + \rho T_0 - \rho T_1)^T A T_0 - k_1(X_0^T A T_1) + k_0 k_1 (1 - T_0^T A T_1) \\ = \; & -\rho(X_0^T A T_1) + k_0 \rho (1 - T_0^T A T_1) - k_1 (X_0^T A T_1) + k_0 k_1 (1 - T_0^T A T_1) \\ = \; & (k_1 + \rho)[k_0(1 - T_0^T A T_1) - (X_0^T A T_1)] \\ = \; & (k_1 + \rho)[k_0(1 - T_0^T A T_1) - (X_1 + \rho T_1 - \rho T_0)^T A T_1] \\ = \; & (k_1 + \rho)[k_0(1 - T_0^T A T_1) - \rho(1 - T_0^T A T_1)] \\ = \; & (1 - T_0^T A T_1)(k_0 - \rho)(k_1 + \rho). \end{aligned}$$

Here $1 - T_0^T A T_1 = \frac{1}{2}(T_0 - T_1)^T A (T_0 - T_1) \neq 0$ since $T_0 \neq T_1$. So Eqn.(5.10) factors, with solutions $k_0 = \rho$ or $k_1 = -\rho$. First we consider the solutions $(k_0, k_1) = (\rho, k_1)$, where $k_1 \neq -\rho$. Here we can assume that $k_1 \neq -\rho$, for otherwise $k_0 = \rho$ and $k_1 = -\rho$ would result in

$$Y_0 = X_0 + k_0 T_0 = X_0 + \rho T_0 = X_1 + \rho T_1 = X_1 - k_1 T_1 = Y_1,$$

which has been excluded by the assumption of the lemma. Therefore by (5.9),

$$Z = k_1(X_0 + k_0 T_0) + k_0(X_1 - k_1 T_1) = k_1(X_0 + \rho T_0 - \rho T_1) + \rho X_1 = (k_1 + \rho) X_1.$$

Thus $[Z] = [X_1]$ as $k_1 + \rho \neq 0$. Hence the resulting biarc is degenerate. When the solutions $(k_0, -\rho)$ are considered, where $k_0 \neq \rho$, it can similarly be shown that $[Z] = [X_0]$.

(ii) $T_0 = T_1$. In this case $X_0^T AT_1 = X_0^T AT_0 = 0$, $X_1^T AT_0 = X_1^T AT_1 = 0$, and $1 - T_0^T AT_1 = 1 - T_0^T AT_0 = 0$. So Eqn. (5.10), which can be rewritten as

$$\frac{X_0^T AX_1}{k_0 k_1} + \frac{X_1^T AT_0}{k_1} - \frac{X_0^T AT_1}{k_0} + 1 - T_0^T AT_1 = 0,$$

reduces to $(X_0^T AX_1)/(k_0 k_1) = 0$. This equation has solutions $k_0 = \pm\infty$ or $k_1 = \pm\infty$. But when $k_0 = \pm\infty$ and $k_1 = \pm\infty$, we have $[Y_0] = [Y_1] = [T_0]$, which is excluded by the assumption of the lemma. So we assume that one of k_0 and k_1 is finite. First consider the solutions $(k_0, k_1) = (k_0, \infty)$, where k_0 is finite. Then

$$\begin{aligned} [Z] &= [k_1(X_0 + k_0 T_0) + k_0(X_1 - k_1 T_1)] = [k_1(X_0 + k_0 T_0 - k_0 T_1) + k_0 X_1] \\ &= [k_1 X_0 + k_0 X_1] = [X_0]. \end{aligned}$$

Hence the resulting biarc is degenerate. The cases in which k_0 and k_1 take other combinations of the allowable values can be proved similarly. □

Based on the foregoing discussion we have proved the following complete existence theorem for proper spherical biarcs interpolating singular data.

Theorem 5.3.12: *Let D be any singular data on S^{d-1}. Then there are ∞^{d-2} many proper spherical biarcs interpolating D and these biarcs are given in Theorem 5.3.10.*

We note that when D is regular the family of biarcs for D has one degree of freedom, which is independent of the dimension of S^{d-1}; when D is singular the family of biarcs for D has $d-2$ degrees of freedom. Thus when $d > 3$, there are more proper spherical biarcs interpolating singular data than those interpolating regular data. When $d = 3$ the number of degrees of freedom for both cases is one.

5.4 Properties of spherical biarcs

5.4.1 Fundamental properties

It has been shown in Theorem 5.3.10 that there are ∞^1 many proper biarcs interpolating any regular data on S^{d-1}. In this section we shall investigate more properties of spherical biarcs. Among them, the most important is that the locus of the joints of all proper spherical biarcs interpolating any regular data D is a circle on S^{d-1} passing through X_0 and X_1. As only the joints of proper biarcs are relevant, of course, the points X_0 and X_1 should be excluded from the locus.

Theorem 5.4.1: *Let $D = \{X_0, T_0, X_1, T_1\}$ be regular data on $S^{d-1} \subset E^d$. Then the joints of all proper biarcs for D form a circle on S^{d-1} passing through X_0 and X_1, but excluding X_0 and X_1.*

We need the following lemma to prove the theorem.

Lemma 5.4.2: *For regular data $D = \{X_0, T_0, X_1, T_1\}$, Eqn. (5.10) does not factor, i.e. it is irreducible.*

PROOF: The discriminant of a bilinear equation $axy + bx + cy + d = 0$, $a \neq 0$, is defined to be $ad - bc$. Eqn. (5.10) is irreducible if and only if its discriminant does not vanish. So we will show that its discriminant

$$\begin{aligned} h &= (X_0^T A X_1)(1 - T_0^T A T_1) + (X_1^T A T_0)(X_0^T A T_1) \\ &= (X_0^T A X_1) + (X_1^T A T_0)(X_0^T A T_1) - (X_0^T A X_1)(T_0^T A T_1) \qquad (5.11) \end{aligned}$$

is nonzero when D is regular. First we consider the case where D is noncoplanar, i.e. X_0, T_0, X_1 and T_1 are linearly independent, and then the case where D is coplanar but regular.

(i) D is noncoplanar. Let $L = [X_0\, T_0\, X_1\, T_1]$, which is a $(d+1) \times 4$ matrix. Since D is noncoplanar, L has full column rank. Let $M = L^T A L$, where $A = \begin{bmatrix} I_d & 0 \\ 0 & -1 \end{bmatrix}$. Then

$$\begin{aligned} M &= L^T A L = \begin{bmatrix} X_0^T \\ T_0^T \\ X_1^T \\ T_1^T \end{bmatrix} A\, [X_0\, T_0\, X_1\, T_1] \\ &= \begin{bmatrix} 0 & 0 & X_0^T A X_1 & X_0^T A T_1 \\ 0 & 1 & X_1^T A T_0 & T_0^T A T_1 \\ X_0^T A X_1 & X_1^T A T_0 & 0 & 0 \\ X_0^T A T_1 & T_0^T A T_1 & 0 & 1 \end{bmatrix}. \end{aligned}$$

By expanding the determinant $\det(M)$ about the first two rows, it is easily verified that

$$\begin{aligned} \det(M) &= [(X_1^T A T_0)(X_0^T A T_1) - (X_0^T A X_1)(T_0^T A T_1)]^2 - (X_0^T A X_1)^2 \\ &= h[(X_1^T A T_0)(X_0^T A T_1) - (X_0^T A X_1)(1 + T_0^T A T_1)]. \qquad (5.12) \end{aligned}$$

We now want to show that $\det(M) \neq 0$. To see this, let us consider the quadratic equation $U^T M U = 0$, where U is a 4 by 1 vector. Geometrically, this quadratic equation stands for the 2-dimensional sphere which is the intersection of S^{d-1} with the 3-dimensional affine manifold determined by X_0, T_0, X_1, T_1. In E^d this intersection is either empty, or a single point, or a proper real sphere. Since it contains at least two distinct points X_0 and X_1 on S^{d-1}, it is a proper 2-dimensional real sphere. So its coefficient matrix M is nonsingular. Hence $\det(M) \neq 0$.

Because $\det(M) \neq 0$, we obtain $h \neq 0$. Incidentally, observe that the other factor of $\det(M)$ in (5.12) is the discriminant of the equation of k_0 and k_1 resulting from (5.5).

(ii) D is coplanar. From the classification of coplanar data following Definition 5.3.2, and the fact that D is regular, we have $X_0 + \rho T_0 = X_1 - \rho T_1$ for some $\rho \neq 0$, or $T_0 = -T_1$. In the former case,

$$\begin{aligned} h &= (X_0^T A X_1)(1 - T_0^T A T_1) + (X_1^T A T_0)(X_0^T A T_1) \\ &= [X_0^T A(X_0 + \rho T_0 + \rho T_1)](1 - T_0^T A T_1) + [(X_0 + \rho T_0 + \rho T_1)^T A T_0](X_0^T A T_1) \\ &= \rho(X_0^T A T_1)(1 - T_0^T A T_1) + \rho(1 + T_0^T A T_1)(X_0^T A T_1) \\ &= 2\rho(X_0^T A T_1) \\ &= 2X_0^T A(X_0 + \rho T_0 + \rho T_1) \\ &= 2X_0^T A X_1 \neq 0. \end{aligned}$$

In the latter case, since $X_0^T A T_1 = -X_0^T A T_0 = 0$, $X_1^T A T_0 = -X_1^T A T_1 = 0$, and $T_0^T A T_1 = -1$, we have

$$h = (X_0^T A X_1)(1 - T_0^T A T_1) + (X_1^T A T_0)(X_0^T A T_1) = 2X_0^T A X_1 \neq 0.$$

□

Proof of Theorem 5.4.1: By Theorem 5.3.9, we just have to show that the locus of Z given by (5.9) is a circle on S^{d-1} passing through X_0 and X_1. By (5.9) and (5.3), Z is given by

$$Z = k_1 Y_0 + k_0 Y_1 = k_0 X_1 + k_1(X_0 + k_0 T_0 - k_0 T_1), \tag{5.13}$$

where k_0 and k_1 satisfy (5.10). By Lemma 5.4.2, (5.10) is irreducible; therefore k_1 can be expressed in terms of k_0, that is,

$$k_1 = \frac{X_0^T A X_1 + k_0 X_1^T A T_0}{X_0^T A T_1 + k_0(T_0^T A T_1 - 1)}.$$

Substituting this expression into (5.13) and omitting a nonzero multiplicative factor, we have

$$Z(k_0) \;=\; k_0[X_0^T A T_1 + k_0(T_0^T A T_1 - 1)]X_1$$

$$\begin{aligned} & +(X_0^T A X_1 + k_0 X_1^T A T_0)(X_0 + k_0 T_0 - k_0 T_1) \\ = \; & k_0^2[(T_0^T A T_1 - 1)X_1 + (X_1^T A T_0)(T_0 - T_1)] \\ & + k_0[(X_0^T A T_1)X_1 + X_0^T A X_1 (T_0 - T_1) + (X_1^T A T_0)X_0] \\ & + (X_0^T A X_1)X_0. \end{aligned} \tag{5.14}$$

Let us consider the coefficient of k_0^2. Since D is regular, $T_0 \neq T_1$. So $T_0^T A T_1 - 1 = -\frac{1}{2}(T_0 - T_1)^T A (T_0 - T_1) \neq 0$. Observing that the last component of $(T_0 - T_1)$ is zero, we conclude that the coefficient of k_0^2 in (5.14) is a nonzero vector. So (5.14) is a rational quadratic curve on S^{d-1}.

Now we need to show that the curve (5.14) is proper, i.e. it is a circle on S^{d-1}. We will prove this using a geometric argument. The curve (5.14) can be a single point, part of a straight line or a circle. Since $Z(k_0)|_{k_0=0} = (X_0^T A X_1)X_0$, the curve passes through X_0; symmetrically, it also passes through X_1. So the curve cannot degenerate into a point. Also it cannot be part of a straight line since the sphere S^{d-1} does not contain the line segment $\overline{X_0 X_1}$. Hence (5.14) must be a circle on S^{d-1}.□

For some applications one may hope to accomplish the interpolation of a data set using a curve with reasonably small windings. For spherical biarc interpolation, this requirement leads to the following question: Is it possible to use spherical biarcs consisting of only minor arcs to interpolate any data $D = \{X_0, T_0, X_1, T_1\}$? The following argument shows that this is possible for regular data but impossible for singular data. Here a minor arc is assumed to be less than a semicircle.

Obviously, from the proof of Theorem 5.3.10, no such solutions exist for singular data, for the two arcs of the interpolating biarc always have to take weights of opposite signs. For regular data D both arcs of a biarc are minor arcs if and only if $k_0 > 0$ and $k_1 > 0$, where k_0 and k_1 are finite solutions of Eqn. (5.10). The next theorem asserts that for regular data there always exist such solutions of Eqn. (5.10).

Theorem 5.4.3: *Eqn. (5.10) has solutions $k_0 > 0$ and $k_1 > 0$ for any regular data $D = \{X_0, T_0, X_1, T_1\}$. That is, for any regular data D there exist spherical biarcs interpolating D that consist of minor arcs only.*

PROOF: Let

$$F(k_0, k_1) \equiv X_0^T A X_1 + k_0(X_1^T A T_0) - k_1(X_0^T A T_1) + k_0 k_1 (1 - T_0^T A T_1). \tag{5.15}$$

Then Eqn. (5.10) is just $F(k_0, k_1) = 0$. First observe that

$$F(0,0) = X_0^T A X_1 = -\frac{1}{2}(X_0 - X_1)^T A (X_0 - X_1) < 0.$$

Since D is regular, $T_0 \neq T_1$. So the coefficient of k_0k_1 in (5.15) is

$$1 - T_0^T A T_1 = \frac{1}{2}(T_0 - T_1)^T(T_0 - T_1) > 0.$$

Therefore $F(k_0, k_1) > 0$ for sufficiently large positive k_0 and k_1. Hence there exist $k_0 > 0$ and $k_1 > 0$ such that $F(k_0, k_1) = 0$. □

From the above proof it follows, geometrically, that the straight line $k_0 = k_1$ always intersects the hyperbola represented by (5.15) in the first quadrant, i.e. where $k_0 > 0$ and $k_1 > 0$. Let $k_0 = k_1$ in (5.15). Then a particular positive solution of Eqn. (5.10) is given by

$$k_0 = k_1 = \frac{-b + \sqrt{b^2 - 4ac}}{2a} \tag{5.16}$$

where

$$a = 1 - T_0^T A T_1, \quad b = X_1^T A T_0 - X_0^T A T_1, \quad c = X_0^T A X_1.$$

By Lemma 5.4.2, when D is regular, (5.10) represents a proper hyperbola with asymptotes parallel to the two axes of the k_0-k_1 plane. Therefore, in general, k_0 does not depend on k_1 continuously throughout the interval $(-\infty, \infty)$. But, to adjust the shape of a biarc by changing the two parameters k_0 and k_1, continuous dependence between the two parameters in a certain range is desirable. The next theorem shows that this is always possible when k_0 and k_1 are both positive.

Theorem 5.4.4: *For regular data D, all solutions (k_0, k_1) of (5.10), where $k_0 > 0$, $k_1 > 0$, are related by a continuous rational linear function $k_1 = f_1(k_0)$, defined on $0 \leq \alpha_0 < k_0 < \alpha_1 \leq \infty$ for constants α_0 and α_1; symmetrically, k_0 can be expressed in terms of k_1 in the same manner.*

PROOF: Obviously, a rational linear function $k_1 = f_1(k_0)$ can be obtained from (5.10), which is of the form $F(k_0, k_1) = 0$. We just need to show that only one branch of the hyperbola $F(k_0, k_1) = 0$ intersects the first quadrant of the k_0-k_1 plane.

First we show that the hyperbola $F(k_0, k_1) = 0$ does not intersect the straight line $k_0 + k_1 = 0$ in the k_0-k_1 plane. This is true if $F(k_0, -k_0) = 0$ does not have real solutions. This can be proved by showing that the discriminant h of the quadratic equation $F(k_0, -k_0) = 0$ is negative. In fact,

$$\begin{aligned} h &= (X_1^T A T_0 + X_0^T A T_1)^2 - 4(X_0^T A X_1)(T_0^T A T_1 - 1) \\ &= [(X_0 - X_1)^T A (T_0 - T_1)]^2 - [(X_0 - X_1)^T A (X_0 - X_1)][(T_0 - T_1)^T A (T_0 - T_1)] \\ &= [(X_0 - X_1)^T (T_0 - T_1)]^2 - [(X_0 - X_1)^T (X_0 - X_1)][(T_0 - T_1)^T (T_0 - T_1)] \\ &< 0. \end{aligned}$$

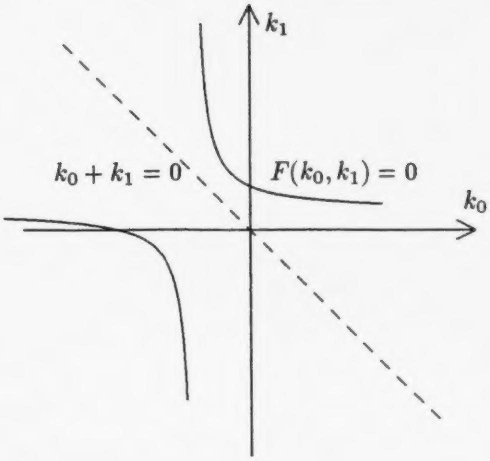


Figure 5.4.1 The hyerpola $F(k_0, k_1) = 0$.

The last strict inequality follows from the Cauchy-Schwarz inequality, and the fact that $X_0 - X_1 \neq \rho(T_0 - T_1)$ for any ρ since D is regular.

Because the hyperbola $F(k_0, k_1) = 0$ does not intersect the straight line $k_0 + k_1 = 0$, it is the translation of a hyperbola $k_0 k_1 = r$, $r > 0$, as shown in Fig. 5.4.1. In the proof of Theorem 5.4.3, it is shown that $F(0, 0) < 0$ and $F(k_0, k_1) > 0$ for sufficiently large positive k_0 and k_1. Thus only one branch of the the hyperbola intersects the first quadrant. □

In fact, (5.10) can be expressed as

$$(k_0 - k_0^c)(k_1 - k_1^c) = \Delta/(1 - T_0^T A T_1)^2,$$

where (k_0^c, k_1^c), the center of the hyperbola, is given by

$$k_0^c = \frac{X_0^T A T_1}{1 - T_0^T A T_1}, \qquad k_1^c = \frac{X_1^T A T_0}{T_0^T A T_1 - 1}, \tag{5.17}$$

and

$$\Delta = (X_0^T A X_1)(T_0^T A T_1 - 1) - (X_0^T A T_1)(X_1^T A T_0).$$

Since the hyperbola $F(k_0, k_1) = 0$ is a translation of hyperbola $k_0 k_1 = r > 0$, $\Delta/(1 - T_0^T A T_1)^2 > 0$. It follows that $\Delta > 0$ since $1 - T_0^T A T_1 > 0$ for regular data.

The range (α_0, α_1) referred to in Theorem 5.4.4 can be determined as follows. The center of the hyperbola $F(k_0, k_1) = 0$ is given by (5.17). So we have the following cases:

1. $\alpha_0 = k_0^c$, $\alpha_1 = \infty$ when $k_0^c \geq 0$, $k_1^c \geq 0$;
2. $\alpha_0 = k_0^c$, $\alpha_1 = -(X_0^T A X_1)/(X_1^T A T_0)$ when $k_0^c \geq 0$, $k_1^c < 0$;
3. $\alpha_0 = 0$, $\alpha_1 = \infty$ when $k_0^c < 0$, $k_1^c \geq 0$;
4. $\alpha_0 = 0$, $\alpha_1 = -(X_0^T A X_1)/(X_1^T A T_0)$ when $k_0^c < 0$, $k_1^c < 0$.

In the above α_1 is found by setting $k_1 = 0$ in (5.10) and solving for k_0.

The next property concerns the reproduction of a circular arc on S^{d-1} by the spherical biarc.

Theorem 5.4.5: *Let the regular data $D = \{X_0, T_0, X_1, T_1\}$ be interpolated by a circular arc C on S^{d-1}. Then every biarc interpolating D either coincides with C or covers the circle that contains C. Moreover, every biarc interpolating D that consists of only minor arcs coincides with C.*

PROOF: Let P be the plane on which C lies. Then X_0, T_0, X_1 and T_1 are coplanar and lie on P. From Theorem 5.4.1 and (5.14), the locus of the joint Z is a circle C' on S^{d-1} in the affine manifold spanned by X_0, T_0, X_1 and T_1, which in this case is P. Thus C' is the circle on S^{d-1} that contains C. Then a biarc coincides with C when $Z \in C - \{X_0, X_1\}$ or covers C' including C twice when $Z \in C' - C$. Obviously, a biarc for D consisting of only minor arcs cannot cover C', so it reproduces C. □

Note that the data that can be interpolated by a single arc is regular coplanar data.

Now we prove one more property of spherical biarcs. This property states that, for generic data $D = \{X_0, T_0, X_1, T_1\}$ and a generic point K on S^2, there is a unique proper spherical biarc interpolating D on S^2 and passing through K.

Lemma 5.4.6: *Let $D = \{X_0, T_0, X_1, T_1\}$ be noncoplanar regular or singular data on S^{d-1}. Let B and B' be two different spherical biarcs interpolating D on S^{d-1}. Then B and B' do not intersect except at X_0 and X_1.*

PROOF: First consider the case where D is noncoplanar regular data. Let F be the 3-*d* affine manifold determined by D. Then all the spherical biarcs interpolating D are on the two dimensional sphere $\bar{S} \equiv S^{d-1} \bigcap F$. Let J be the circle which is the locus of joints of all biarcs interpolating D. Obviously $J \subset \bar{S}$, and J divides $\bar{S}$ into two disjoint parts, denoted by $\bar{S}_0$ and $\bar{S}_1$.

Let the biarc B be composed of two arcs C_0 and C_1, and B' composed of two arcs C_0' and C_1'. Let Z, $Z' \in J$ be the joints of biarcs B and B'. Obviously, a proper spherical biarc interpolating noncoplanar regular data can not be contained in the circle J, for otherwise the data would be coplanar. Therefore we can assume that C_0 and C_0' are in $\bar{S}_0$, and C_1 and C_1' are in $\bar{S}_1$. Since B and B' are different biarcs, Z and Z' are different points, for the spherical biarcs interpolating D are in one-to-one correspondence with the points on J. Therefore arcs C_0 and C_0' do not intersect in $\bar{S}_0$, except at X_0, since they have the common tangent T_0 at X_0. Similarly, arcs C_1 and C_1' do not intersect in $\bar{S}_1$, except at X_1. Since C_0 and C_1' belong to different parts $\bar{S}_0$ and $\bar{S}_1$ of $\bar{S}$, they do not intersect. For the same reason C_1 and C_0' do not intersect. Hence the biarcs $B = C_0 \bigcup C_1$ and $B' = C_0' \bigcup C_1'$ do not intersect, except at X_0 and X_1.

When D is singular data, by Theorem 5.3.12, the locus J of joints is a $d-2$ dimensional sphere on S^{d-1}. The sphere J divides S^{d-1} into two parts, still denoted by $\bar{S}_0$ and $\bar{S}_1$. The remaining argument is the same as the one above given for the regular data. □

Theorem 5.4.7: *Let $D = \{X_0, T_0, X_1, T_1\}$ be noncoplanar regular data or singular data on S^{d-1}, $d \geq 3$. Let $\bar{S}$ be the 2-dimensional sphere cut on S^{d-1} by the 3-d affine manifold F determined by D when D is noncoplanar regular data; let $\bar{S}$ be S^{d-1} when D is singular data. For a generic point K on $\bar{S}$, i.e. a point K that is not on any of the degenerate biarcs on $\bar{S}$ interpolating D, there exists a unique proper spherical biarc interpolating D that passes through K.*

Proof: We just need to prove the existence of the spherical biarc described in the theorem; uniqueness follows from Lemma 5.4.6. First consider the case where D is noncoplanar regular data. Let $\bar{S}_0$ and $\bar{S}_1$ be the two parts into which $\bar{S}$ is divided by the circle J, the locus of joints of biarcs interpolating D. Then for a generic point $K \neq X_0$ on $\bar{S}_0 \subset \bar{S}$, the 2-dimensional plane P determined by the tangent line of $\bar{S}$ at X_0 with direction T_0 and the point K cuts $\bar{S}$ in a circle C.

Now we claim that T_0 is not tangent to the circle J at X_0. To see this, let us consider a proper biarc B interpolating D with the joint $Z' \in J$. Let C_0 and C_1 be the two arcs of B such that C_0 starts at X_0 and C_1 ends at X_1. If T_0 is tangent to J at X_0, then C_0 is an arc on J since it is tangent to J at X_0 and intersects J at the joint Z'. It follows then that C_1 must also be an arc on J since C_1 is tangent to J at Z' and intersects J at X_1. Therefore the biarc B is in fact a single arc on J. So D is coplanar regular data. This is a contradiction to our assumption.

Since T_0 is not tangent to J at X_0, the circle C and the circle J, both are on sphere $\bar{S}$, intersect in two points, one of which is X_0, and the other is denoted by Z. We assume that $Z \neq X_0$ and $Z \neq X_1$; this is ensured by assuming that K is not on the circular arc on $\bar{S}_0$ passing through X_0 and X_1 and having tangent T_0 at X_0. Then this joint Z defines a spherical biarc interpolating D and passing through K. When $K \in \bar{S}_1 \subset S$ is not on the circular arc on $\bar{S}_1$ passing through X_0 and X_1 and having tangent T_1 at X_1, the conclusion can be proved similarly by considering the circle determined by the point X_0, the tangent T_1 of $\bar{S}$ at X_1, and the point K.

Now consider the case of singular data D. Again $\bar{S} = S^{d-1}$ is divided into two disjoint parts $\bar{S}_0$ and $\bar{S}_1$ by the locus J of joints, which is a $d-2$ dimensional sphere on S^{d-1}. Now the remaining proof is similar to the one above given for noncoplanar regular data. □

Theorem 5.4.7 is illustrated in Fig. 5.3.3. We remark that there are two degenerate biarcs interpolating D when D is noncoplanar regular data, and there is only one such biarc when D is singular. The case where D is coplanar regular data is not governed by this theorem. For coplanar regular data, all interpolating spherical biarcs are contained in a circle, as indicated by Theorem 5.4.5.

Another simple property of spherical biarcs is that a spherical biarc interpolating D does not intersect with itself if D is noncoplanar regular data; when D is coplanar and regular, a spherical biarc interpolating D might overlap with itself as its two arcs are contained in the same circle.

5.4.2 Perturbation of singular data

In Section 5.3 we saw that the spherical biarc solutions for regular and singular data have different structures. For data D on $S^{d-1} \subset E^d$, the locus of joints of all spherical biarcs interpolating D is a circle when D is regular; the locus is a $d-2$ dimensional sphere on S^{d-1} when D is singular. On the other hand, regular data is generic and a small perturbation to singular data can make it regular. Therefore we ask the following question: Let D' be regular data obtained from singular data D through a small perturbation α to D. Let the loci of joints for D' and D be J' and J respectively. Is the circle J' close to the sphere J? The significance of this problem is as follows: using the above notation, if we have a spherical biarc C' interpolating regular data D', can we expect to find a biarc C interpolating singular data D so that C is very close to C'?

The answer to this question is yes. So, as the perturbation yielding D' tends to zero, the circle J' approaches the sphere J, and finally merges into it. Because the perturbation α can assume different directions, the corresponding circle J' approaches J from different directions.

We shall consider a special case in the next theorem, that is, only the tangent direction T_0 in D is subjected to a perturbation. But a similar argument applies to the case where the points X_0 and X_1 and the tangents T_0 and T_1 are all subjected to perturbations. When only T_0 is subjected to a perturbation α, the new data is $D' = \{X_0, T_0 + \alpha, X_1, T_1\}$. By convention, the perturbation α must satisfy certain conditions.

Definition 5.4.1: *Let $\alpha \neq 0$ be a point at infinity in E^d. Then α is called an* allowable perturbation *if*

$$X_0^T A(T_0 + \alpha) = 0, \quad (T_0 + \alpha)^T A(T_0 + \alpha) = 1. \tag{5.18}$$

Theorem 5.4.8: *Let $D = \{X_0, T_0, X_1, T_1\}$ be singular data on $S^{d-2} \subset E^d$, $d \geq 3$. Let $D' = \{X_0, T_0 + \alpha, X_1, T_1\}$ be regular data on S^{d-1}, where α is a sufficiently small allowable perturbation. Let J and J' be the loci of joints of spherical biarcs for D and D' respectively. Then the circle J' approaches the sphere J uniformly as $|\alpha| = \sqrt{\alpha^T \alpha} \to 0$.*

First we need some lemmas.

Lemma 5.4.9: *Let $D = \{X_0, T_0, X_1, T_1\}$ be singular data on S^{d-1}. For a sufficiently small allowable perturbation $\alpha \neq 0$, the points X_0, $T_0 + \alpha$, X_1 and T_1 are linearly independent.*

PROOF: The proof is by contradiction. If they are linearly dependent, since X_0 and X_1 are in normalized form, and T_0 and T_1 are points at infinity, then we have two possible cases. (i) $T_0 + \alpha = \sigma T_1$ for some σ; or (ii) $X_0 + \rho_0(T_0 + \alpha) = X_1 + \rho_1 T_1$ for some ρ_0 and ρ_1.

In case (i) we have

$$(T_0 + \alpha)^T A(T_0 + \alpha) = \sigma^2 T_1^T A T_1.$$

So $\sigma^2 = 1$. When $\sigma = 1$, we have $T_0 + \alpha = T_1$, or $\alpha = T_1 - T_0$. This contradicts the assumption that α is sufficiently small and $\alpha \neq 0$, for $T_1 - T_0$ is constant. When $\sigma = -1$, a contradiction follows similarly.

In case (ii) we have

$$[X_0 + \rho_0(T_0 + \alpha)]^T A[X_0 + \rho_0(T_0 + \alpha)] = (X_1 + \rho_1 T_1)^T A(X_1 + \rho_1 T_1).$$

Simplifying this equation using (5.18), we obtain $\rho_0^2 = \rho_1^2$. When $\rho_0 = \rho_1$, we have

$$X_0 + \rho_0(T_0 + \alpha) = X_1 + \rho_0 T_1. \tag{5.19}$$

Clearly $\rho_0 \neq 0$ since $X_0 \neq X_1$. So

$$\alpha = \rho_0^{-1}(X_1 - X_0) + (T_1 - T_0).$$

Multiplying $X_0^T A$ to both sides of (5.19), and by Definition 5.4.1, we find

$$X_0^T A X_1 + \rho_0 X_0^T A T_1 = 0,$$

that is, $\rho_0^{-1} = -X_0^T A T_1 / X_0^T A X_1$. Substituting this in the above expression for α,

$$\alpha = -\frac{X_0^T A T_1}{X_0^T A X_1}(X_1 - X_0) + (T_1 - T_0).$$

This is contradictory to that α is sufficiently small and $\alpha \neq 0$. When $\rho_0 = -\rho_1$, a contradiction follows similarly. □

Lemma 5.4.10: *Let* $U = (X_0 - X_1)/|X_0 - X_1|$ *and* $V = \alpha/|\alpha|$ *be two vectors of unit length in* E^d. *Let* γ *be the angle between* U *and* V, $0 \leq \gamma \leq \pi$. *Then there exists a constant* $c > 0$, *such that* $\sin\gamma \geq c$.

PROOF: Since $X_0^T A(T_0 + \alpha) = 0$ and $X_0^T A T_0 = 0$, we have $X_0^T A\alpha = 0$, or $X_0^T A V = 0$. So the direction V is parallel to the hyperplane $X_0^T A X = 0$.

On the other hand, as $X_0^T A X_0 = 0$ and $X_0^T A X_1 \neq 0$, the direction $X_0 - X_1$ is not parallel to the hyperplane $X_0^T A X = 0$. Let β be the angle formed between the direction $X_0 - X_1$ and the hyperplane $X_0^T A X = 0$. Then we can assume $0 < \beta \leq \pi/2$. Since, by definition, β is the minimum of angles formed between direction $X_0 - X_1$ and all the directions parallel to hyperplane $X_0^T A X = 0$, we have $\beta \leq \gamma \leq \pi - \beta$. Therefore $\sin\gamma \geq \sin\beta$. Setting $c = \sin\beta$ completes the proof. □

Lemma 5.4.11: *Let* $S \subset E^3$ *be a sphere of radius* r. *Let* J_0 *and* J_1 *be two circles on* S *intersecting at two distinct points. Let the Hausdorff distance of* J_0 *and* J_1 *be defined as*

$$d_H(J_0, J_1) = \max_x\{\min_y\{d(x,y) | x \in J_0, y \in J_1\}, \min_y\{d(y,x) | y \in J_0, x \in J_1\}\},$$

where $d(x,y)$ *is the Euclidean distance between points* x *and* y. *Then*

$$d_H(J_0, J_1) \leq 2r\sin\theta,$$

where θ *is the angle between the two planes containing* J_0 *and* J_1, *respectively, and it is assumed that* $0 \leq \theta \leq \pi/2$.

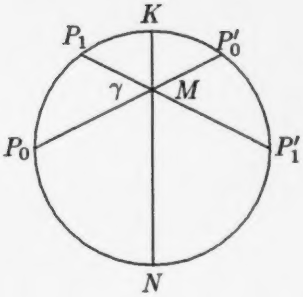


Figure 5.4.2 The section of the sphere S intersected by the plane KNP_0.

PROOF: When $\theta = 0$, the lemma is trivially true. So assume that $0 < \theta \leq \pi/2$. Let J_0 and J_1 intersect at points X_0, $X_1 \in S$, where $X_0 \neq X_1$. In the following we first show that for any point $P_0 \in J_0$,

$$\min_y \{d(P_0, y) \mid y \in J_1\} \leq 2r \sin\theta. \tag{5.20}$$

Let J_0 and J_1 be contained in the two planes q_0 and q_1, respectively. Let q be the plane that passes through the straight line X_0X_1 and bisects the angle θ between q_0 and q_1. Let points K, $N \in S$ be such that segment $\overline{KN}$ passes through the midpoint M of $\overline{X_0X_1}$ and is perpendicular to plane q. Note the segment $\overline{KN}$ is unique. For any point $P_0 \in J_0$ with $P_0 \neq X_0$ and $P_0 \neq X_1$, let KNP_0 denote the unique plane containing K, N and P_0; when $P_0 = X_0$ or $P_0 = X_1$, (5.20) is trivially true because $\min_y\{d(P_0, y) \mid y \in J_1\} = 0$.

Let C be the circle intersected on S by the plane KNP_0. Suppose that KNP_0 intersects the plane q_0 of J_0 and the plane q_1 of J_1 in two lines P_0P_0' and P_1P_1', respectively, where P_0', P_1 and P_1' are on the circle C. The section of S intersected by the plane KNP_0 is illustrated in Fig. 5.4.2.

Without loss of generality, assume that $\angle P_0MP_1 \leq \angle P_1MP_0'$. Let $\gamma = \angle P_0MP_1$. Then by the above construction it follows that $\gamma \leq \theta$; in fact, equality holds if and only the plane KNP_0 is perpendicular to the line X_0X_1. From elementary plane geometry it is known that an angle inside a circle is equal to half the sum of the degrees of the two arcs of the circle subtended by the angle and its opposite angle. Let $\text{arc}(P_0P_1)$ be the length of the minor arc of C between P_0 and P_1, and $\text{arc}(P_0'P_1')$ be defined similarly. Then we have

$$\text{arc}(P_0P_1) + \text{arc}(P_0'P_1') = 2r'\gamma,$$

where r' is the radius of C. Since $r' \leq r$, it follows that

$$\mathrm{arc}(P_0P_1) + \mathrm{arc}(P_0'P_1') \leq 2r\gamma.$$

Thus

$$\mathrm{arc}(P_0P_1) \leq 2r\gamma.$$

Therefore

$$\min_y \{d(P_0, y) |\ y \in J_1\} \leq |\overline{P_0P_1}| \leq 2r \sin\gamma \leq 2r \sin\theta,$$

where $|\overline{P_0P_1}|$ is the length of the segment $\overline{P_0P_1}$. Since P_0 is an arbitrary point on J_0, we have

$$\max_x \{\min_y \{d(x,y) | x \in J_0, y \in J_1\} \leq \max_{P_0 \in J_0} \{|\overline{P_0P_1}|\} \leq 2r \sin\theta.$$

Symmetrically, we have

$$\max_x \min_y \{d(y,x) | y \in J_0, x \in J_1\} \leq 2r \sin\theta.$$

Combining these two inequalities completes the proof. □

PROOF of Theorem 5.4.8: To determine the distance from the circle J' to the sphere J, we consider the follow measurement

$$D(J', J) = \max_x \min_y \{d(x,y) | x \in J',\ y \in J\},$$

where $d(x, y)$ is the Euclidean distance between point x and y. Note that the above measurement is not symmetric and is different from the Hausdorff distance. Our goal is to prove that $D(J', J) \to 0$ uniformly for all α as $|\alpha| \to 0$.

Recall that J, as the locus of joints for singular data D, is a $d-2$ dimensional sphere on S^{d-1}. The idea of the following proof is to intersect S^{d-1} with an appropriate 3-d affine manifold $\mathcal{F}$ so that $J' \subset \mathcal{F}$ and that $\bar{J} = J \bigcap \mathcal{F}$ is a circle. Then we can estimate the distance between the two circles J' and $\bar{J}$ with respect to some perturbation α, which is easier than estimating the distance from the circle J' to the sphere J directly.

Since α is sufficiently small, by Lemma 5.4.9, we can assume that X_0, $T_0 + \alpha$, X_1 and T_1 are linearly independent. Let $\mathcal{F}$ be the 3-d affine manifold determined by these four points. Then $\bar{S}^2 = S^{d-1} \bigcap \mathcal{F}$ is a 2-dimensional sphere on S^{d-1}. Now introduce a coordinate system $\bar{E}^3$ in the 3-d Euclidean space $\mathcal{F}$ with the origin $\bar{O}$ at the center of $\bar{S}^2$. Let the equation of $\bar{S}^2$ in $\bar{E}^3$ be $X^T \bar{A} X = 0$, where $\bar{A} = \mathrm{diag}[1, 1, 1, -r^2]$ and r is the radius of $\bar{S}^2$. Clearly, D and D' are still singular and regular data on $\bar{S}^2$. For simplicity, we use the same symbols to denote D, D' and the points they are composed of in $\bar{E}^3$. Now we have two spherical biarc problems on $\bar{S}^2$ with data D and D', respectively. Let $\bar{J} = J \bigcap \mathcal{F}$ and $\bar{J}' = J'$. Clearly, by the construction of $\mathcal{F}$, $\bar{J}$ and $\bar{J}'$ are the loci of joints for D and D' in $\bar{S}^2$, and both $\bar{J}$ and $\bar{J}'$ are circles passing through X_0 and X_1 on $\bar{S}^2$.

Below we shall be concerned with the geometry of $\bar{E}^3$ when we estimate the distance between the circles $\bar{J}'$ and $\bar{J}$. First consider the case where D is singular with $X_0 + \rho T_0 = X_1 + \rho T_1$ for some $\rho \neq 0$. In this case the unit normal to the plane containing $\bar{J}$ is, by Theorem 5.3.10,

$$N = \frac{Y - \bar{O}}{|Y - \bar{O}|},$$

where $Y = X_0 + \rho T_0 = X_1 + \rho T_1$, and $\bar{O} = [0,0,0,1]^T$ is the center of $\bar{S}^2$.

Since, by (5.14), the plane containing $\bar{J}'$ is spanned by the points X_0, X_1 and $T_0 + \alpha - T_1$, a normal vector to this plane is

$$\begin{aligned} (X_0 - X_1) \times (T_0 + \alpha - T_1) &= (X_0 - X_1) \times [(T_0 - T_1) + \alpha] \\ &= (X_0 - X_1) \times [\rho^{-1}(X_1 - X_0) + \alpha] \\ &= (X_0 - X_1) \times \alpha. \end{aligned}$$

In the above simplification only the linearity of the cross product and the fact that the cross product of a vector with itself is the zero vector are used. Since the points are in homogeneous representation, the actual manipulation of the cross product based on 3-d vector notation is illegal here.

Let θ be the angle between the two planes containing J and J'. Then θ is also the angle between the normal vectors of the two planes. Setting $V = \alpha/|\alpha|$, we have

$$\begin{aligned} \sin\theta &= \frac{|[(X_0 - X_1) \times \alpha] \times (Y - \bar{O})|}{|(X_0 - X_1) \times \alpha|\,|Y - \bar{O}|} \\ &= \frac{|[(X_0 - X_1) \times V] \times (Y - \bar{O})|}{|(X_0 - X_1) \times V|\,|Y - \bar{O}|} \\ &= \frac{\left|[(X_0 - X_1)^T(Y - \bar{O})]V - [V^T(Y - \bar{O})](X_0 - X_1)\right|}{|(X_0 - X_1) \times V|\,|Y - \bar{O}|}. \end{aligned} \tag{5.21}$$

Here the familiar identity

$$(a \times b) \times c = (a \cdot c)b - (b \cdot c)a$$

is used.

Now we will simplify (5.21). First,

$$(X_0 - X_1)^T(Y - \bar{O}) = (X_0 - X_1)^T \bar{A}(Y - \bar{O})$$

$$
\begin{aligned}
&= X_0^T\bar{A}Y - X_1^T\bar{A}Y - (X_0 - X_1)^T\bar{A}\bar{O} \\
&= X_0^T\bar{A}(X_0 + \rho T_0) - X_1^T\bar{A}(X_1 + \rho T_1) \\
&= 0.
\end{aligned}
$$

Secondly, by Definition 5.4.1, we have $X_0^T\bar{A}(T_0 + \alpha) = 0$. Therefore $X_0^T\bar{A}\alpha = 0$, since $X_0^T\bar{A}T_0 = 0$. Hence $X_0^T\bar{A}V = 0$. Finally, also by convention, we have $(T_0+\alpha)^T\bar{A}(T_0+\alpha) = 1$, or

$$T_0^T\bar{A}T_0 + 2T_0^T\bar{A}\alpha + \alpha^T\bar{A}\alpha = 1.$$

Since $T_0^T\bar{A}T_0 = 1$, we have

$$T_0^T\bar{A}\alpha = -\frac{1}{2}\alpha^T\bar{A}\alpha = -\frac{1}{2}\alpha^T\alpha = -\frac{1}{2}|\alpha|^2.$$

So

$$T_0^T\bar{A}V = T_0^T\bar{A}\frac{\alpha}{|\alpha|} = -\frac{1}{2}|\alpha|.$$

By these properties of V, from (5.21) we have

$$
\begin{aligned}
\sin\theta &= \frac{\left|[(X_0 - X_1)^T(Y - \bar{O})]V - [V^T(X_0 + \rho T_0 - \bar{O})](X_0 - X_1)\right|}{|(X_0 - X_1) \times V|\,|Y - \bar{O}|} \\
&= \frac{|\rho|\,|\alpha|\,|X_0 - X_1|}{2|(X_0 - X_1) \times V|\,|Y - \bar{O}|} \\
&= \frac{|\rho|\,|\alpha|}{2|U \times V|\,|Y - \bar{O}|},
\end{aligned}
$$

where $U = (X_0 - X_1)/|X_0 - X_1|$. By Lemma 5.4.10, we have $|U \times V| \geq c$, where $c > 0$ is a constant independent of α. Moreover, since

$$
\begin{aligned}
(Y - \bar{O})^T(Y - \bar{O}) &= (Y - \bar{O})^T\bar{A}(Y - \bar{O}) \\
&= (Y - \bar{O})^T\bar{A}Y - (Y - \bar{O})^T\bar{A}\bar{O} = Y^T\bar{A}Y - \bar{O}^T\bar{A}Y \\
&= (X_0 + \rho T_0)^T\bar{A}(X_0 + \rho T_0) + r^2 = \rho^2 + r^2,
\end{aligned}
$$

we have $|Y - \bar{O}| \geq |\rho|$. So

$$\sin\theta \leq \frac{|\alpha|}{2c}. \tag{5.22}$$

When D is singular with $T_0 = T_1$, let $Y = T_0 = T_1$ in the above proof. We can still prove

$$(X_0 - X_1)^T\bar{A}(Y - \bar{O}) = 0, \;\; X_0^T\bar{A}V = 0,$$

and

$$T_0^T\bar{A}V = -\frac{1}{2}|\alpha|.$$

Using these results, (5.22) can be obtained similarly.

Since $\bar{S}^2$ is on the unit sphere S^{d-1}, its radius $r \leq 1$. By Lemma 5.4.11, we have

$$\max_x\{\min_y\{d(x,y)|x \in \bar{J}',\ y \in \bar{J}\}\} \leq 2r\sin\theta \leq 2\sin\theta.$$

It follows that, since $\bar{J} \subset J$,

$$\begin{aligned} D(J',J) &= \max_x \min_y\{d(x,y)|x \in J',\ y \in J\} \\ &\leq \max_x \min_y\{d(x,y)|x \in J',\ y \in \bar{J}\} \\ &\leq 2\sin\theta \leq \frac{|\alpha|}{c}. \end{aligned}$$

Hence the circle J' tends to the sphere J uniformly as $|\alpha| \to 0$. □

5.5 Summary

In this chapter we have developed the theory of spherical biarc interpolation. Spherical biarcs are used to solve the Hermite interpolation problem on a sphere, i.e. to interpolate two points X_0, X_1 and the associated end tangents T_0 and T_1, collectively called data $D = \{X_0, T_0, X_1, T_1\}$, on a sphere S in E^d, $d \geq 3$. Given such data, an interpolating spherical biarc is determined by the joint, the point where the two arcs meet. It is shown that there are two kinds of data: regular data, which is generic, and singular data. Given regular data, the locus of joints is a circle on the sphere S, while for singular data the locus of joints is a $(d-2)$-dimensional sphere on S.

Several more properties of spherical biarcs are derived. For instance, give any data D on a two dimensional sphere S in E^3, any two different spherical biarcs interpolating D do not intersect each other except at the end points X_0 and X_1. For a general point K on S, there exists a unique spherical biarc interpolating D that passes through K; that is, the tangent vectors of all spherical biarcs interpolating D define a vector field on S, with X_0 and X_1 being the only singular points of the field. The relation between spherical biarcs interpolating regular and singular data is also investigated.

Chapter 6

Shape Control and Applications of Spherical Biarcs

By the shape control of spherical biarcs we mean appropriately choosing the joint of the biarc so that it looks fair. Since for any data $D = \{X_0, T_0, X_1, T_1\}$ on S^{d-1} there exist infinitely many spherical biarcs interpolating D, the determination of a good joint is crucial to the application of spherical biarcs. In the first two sections of this chapter we shall argue, with support from experimental results, that if a simple and explicit formula for determining the joint is desired, the biarc with its two arcs having chords of equal length is an acceptable choice.

In Section 6.3 an algorithm is proposed for interpolating a sequence of points on a sphere using spherical biarcs. Biarcs with equal chord lengths are used in this algorithm.

Finally, in Section 6.4, we will address the application of spherical biarcs in orientation interpolation by using a spherical biarc spline to interpolate a point sequence in the unit quaternion space. Compared with other existing solutions to this problem, we show that our solution is simpler and more efficient.

6.1 Criteria and choices for shape control

In the first subsection we will discuss some issues in the shape control of spherical biarcs, from which three possible criteria are derived: the ratio of the two radii of the two arcs, the winding, and the twist of a spherical biarc. Then in the next two subsections we derive two simple formulae to compute the joint of a biarc interpolating regular data D; one gives a biarc with equal chords, the other a biarc with minimal twist. Together with the formula (5.16) which gives $k_0 = k_1 > 0$, we have three formulae available. In the next section these three formulae are evaluated and compared through experiments.

To simplify the terminology, the two radii of the two arcs of a biarc will be called the radii of the biarc, and the smaller one of the two radii when they are not equal will be called the smaller radius of the biarc.

6.1.1 Criteria

There are no fixed mathematical criteria according to which a biarc, or a general interpolating curve, is fair or not. However, in the traditional study of plane biarcs, an empirical criterion is to minimize the difference of the two radii of a biarc, or the difference of the curvatures of the two arcs, or some other similar criteria [Bez72, NuM88, SuL89]. These criteria are used to reduce the jump in the curvature at the joint of the biarc and avoid having one arc with extremely high curvature relative to the other arc, which gives the appearance of a sharp turn on the biarc. We shall regard the ratio of the two radii of a biarc as the most important criterion in our analysis of spherical biarcs.

We have two more concerns in our analysis. The first is that while major circular arcs are useful in constructing fair spherical biarcs, they should be used only when it is necessary, for otherwise excessive winding of the biarc would result. This issue has not been of concern in most previous studies of plane biarcs, in which major circular arcs are not used at all; this usually requires that the data are quite well behaved. For example, even S-shaped plane biarcs are excluded in [SuL89]. Such assumptions are possible and reasonable because of the intended application of plane biarcs; they are mainly used to approximate some existing smooth curves, cubic spline curves say, or to interpolate data points that represent smooth geometric shapes [Bez72, SuL89]. But since a main application of spherical biarcs is to interpolate points in the unit quaternion space which represent some object orientations in E^3, no assumptions on the type of the data should be imposed. This means that in the discussion of spherical biarcs all possible data configurations need to be addressed. In this case, although Theorem 5.4.3 ensures that there always exist proper biarcs consisting of minor arcs interpolating D as long as D is regular, major arcs are still useful because it is hard to find a formula giving a biarc consisting of only minor arcs whose shape is acceptable for any data, especially data that is close to singular data. Recall that when the data D is singular, Theorem 5.3.10 implies that a proper interpolating spherical biarc must contain one major arc. So another criterion is how to use major arcs appropriately to avoid excessive winding in a biarc.

The second concern is specific to space biarcs, which include spherical biarcs. Usually the two arcs of a space biarc are not on the same plane. We are interested in the angle formed by the two planes of the two arcs of the biarc, called *the twist of the biarc*. The twist measures the jump in the torsion of the biarc as a space curve. The problem of minimizing the twist of a space biarc has been discussed in [NuM88], where some restrictions are assumed on the alignment of the data. In the following we want to give a comprehensive analysis of this problem in the setting of spherical biarcs, and see if minimizing the twist can yield a fair spherical biarc.

So far we have introduced three possible criteria to judge the shape of a spherical biarc: (1) minimize the difference of the two radii of a biarc; (2) restrain the use of major arcs in order to avoid excessive winding; (3) minimize the twist of a biarc. These will be called *criterion 1*, *criterion 2* and *criterion 3*, respectively. Since it is usually difficult to achieve all three goals with a single choice of joint, one expects to make compromises among these criteria. In this respect we regard the first two criteria as more important than the third one, unless a certain application requires a minimal twist.

The these three criteria will be incorporated in our analysis as follows. We will initially discuss only regular data, with singular data to be addressed later. Three formulae will be proposed to compute the parameters (k_0, k_1) of the joint of a biarc. The first formula gives a biarc with the chords of its two arcs having equal length, called an *equal chord biarc*; intuitively, such a biarc is reasonably good since the lengths of the two arcs do not differ much when they are both minor arcs. The second formula is given by (5.16), which yields the solution $k_0 = k_1 > 0$; this solution guarantees that the two arcs of the biarc are always minor arcs, thus enforcing criterion 2. This biarc will be referred to as the *equal k biarc*. The third formula gives a spherical biarc with a minimal twist, called a *minimal twist biarc*. The methods for computing the joint by the above three formulae will be called *choice 1*, *choice 2* and *choice 3*, respectively. Because it is not clear whether or not the equation in terms of k_0 and k_1 set up according to criterion 1 can be solved explicitly, instead of using numerical techniques to solve this equation, criterion 1 is used to judge the above three formulae in our investigation as follows. The three biarcs are judged according to which biarc gives the ratio of radii that is closest to one in the worst case and on average. Our experiments, to be explained later in this section, suggest that the equal chords biarc, i.e. choice 1, is the best one among the three.

The ratio of the two radii of a biarc is not the only indicator of its shape. Although a biarc with the two radii not differing too much is considered to be fair, a biarc with the

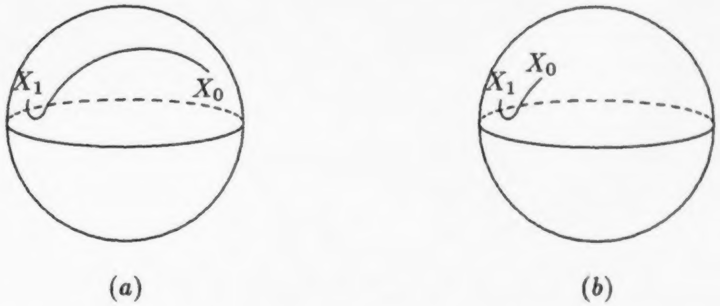


Figure 6.1.1 Comparison of two biarcs: The two biarcs have different shapes, but their two arcs have the same radii.

two radii differing radically may also be acceptable in some cases. For instance, while the biarc shown in Fig. 6.1.1(a) does not look fair because of the sharp turn, the biarc in Fig. 6.1.1(b) should be acceptable; however, these two biarcs have the same radii. They look different because in Fig. 6.1.1(b) the smaller radius of the biarc is not very small relative to the distance $|X_0X_1|$. Therefore, the ratio of the smaller radius of a biarc to the distance $|X_0X_1|$ between two interpolated points should also be taken into account in shape analysis. The significance of this remark is that a biarc cannot be judged as not fair solely because of a large difference between the two radii of its two arcs; only when the radius of one arc of a biarc is considerably small relative to both the radius of the other arc and the distance $|X_0X_1|$ can the biarc be considered to contain a sharp turn.

6.1.2 Equal chord biarc

In this subsection we will give a formula to compute the joint of a spherical biarc with equal chords for its two arcs, and we will study some properties of this biarc. This biarc will be compared with biarcs generated in two other ways later in this section.

Our problem then is that given data $D = \{X_0, T_0, X_1, T_1\}$ on S^{d-1}, how to find a spherical biarc interpolating D such that its two arcs have chords of the same length. The existence of a solution to this problem is obvious by the following geometric observation. A spherical biarc interpolating $D = \{X_0, T_0, X_1, T_1\}$ has equal chords if and only if the joint is on the perpendicular bisecting hyperplane of X_0 and X_1 in E^d, denoted by L. If D is regular, by Theorem 5.4.1, the locus of joints of all spherical biarcs interpolating D is a circle passing through X_0 and X_1, and this circle intersects the hyperplane L in two points.

When D is singular, by Theorem 5.3.12, the locus of joints is a $d-2$ dimensional sphere on S^{d-1}, which intersects the hyperplane L in a $d-3$ dimensional sphere, where $d \geq 3$.

The equation of the hyperplane L is easily seen to be $(X_0 - X_1)^T(2X - X_0 - X_1) = 0$ with X in normalized form. This can be expressed as $(X_0 - X_1)^T A(2X - X_0 - X_1) = 0$. After simplification, using $X_0^T A X_0 = X_1^T A X_1 = 0$, we have

$$(X_0 - X_1)^T A X = 0, \tag{6.1}$$

with X now not necessarily being in normalized form.

First suppose that D is regular. Then substituting $Z(k_0)$ of (5.14) in (6.1) yields

$$\begin{aligned} &k_0^2[(T_0^T A T_1 - 1)(X_0^T A X_1) - (X_1^T A T_0)(X_1^T A T_0 + X_0^T A T_1)] \\ &-2k_0(X_1^T A T_0)(X_0^T A X_1) - (X_0^T A X_1)^2 = 0. \end{aligned}$$

Solving this equation for k_0, we have the two roots

$$k_0^{(1)} = \frac{-X_0^T A X_1}{X_1^T A T_0 - \sqrt{\Delta}}$$

and

$$k_0^{(2)} = \frac{-X_0^T A X_1}{X_1^T A T_0 + \sqrt{\Delta}},$$

where

$$\Delta = (X_0^T A X_1)(T_0^T A T_1 - 1) - (X_0^T A T_1)(X_1^T A T_0).$$

By the remark following the proof of Theorem 5.4.4, $\Delta > 0$ for regular D, thus confirming that the two roots are real and distinct. By (5.10) the two corresponding k_1's are found to be

$$k_1^{(1)} = \frac{-X_0^T A X_1}{-X_0^T A T_1 - \sqrt{\Delta}},$$

$$k_1^{(2)} = \frac{-X_0^T A X_1}{-X_0^T A T_1 + \sqrt{\Delta}}.$$

Hence we have obtained two joints that give equal chord biarcs. These two joints are given by $(k_0^{(1)}, k_1^{(1)})$ and $(k_0^{(2)}, k_1^{(2)})$, respectively.

Although both of these two biarcs have equal chords, we will see that the one given by $(k_0^{(2)}, k_1^{(2)})$ is better, in the sense that it give biarcs consisting of minor arcs in more cases than $(k_0^{(1)}, k_1^{(1)})$. To see this, note that the numerator $-X_0^T A X_1 > 0$ in all the above expressions for $k_0^{(1)}$, $k_1^{(1)}$, $k_0^{(2)}$ and $k_1^{(2)}$. Also

$$X_1^T A T_0 - \sqrt{\Delta} < X_1^T A T_0 + \sqrt{\Delta},$$

and

$$-X_0^T AT_1 - \sqrt{\Delta} < -X_0^T AT_1 + \sqrt{\Delta},$$

i.e. the denominators of $k_0^{(1)}$ and $k_1^{(1)}$ are smaller than those of $k_0^{(2)}$ and $k_1^{(2)}$. It is clear then that at least one of the two biarcs given above consists of two minor arcs if and only if

$$X_1^T AT_0 + \sqrt{\Delta} > 0,$$

and

$$-X_0^T AT_1 + \sqrt{\Delta} > 0,$$

that is, if and only if the biarc given by $k_0^{(2)}$ and $k_1^{(2)}$ consists of two minor arcs. Based on this argument, in the following discussion by the equal chord biarc we mean the one given by $k_0^{(2)}$ and $k_1^{(2)}$.

Experiments showed that the last two inequalities do not hold for some regular data. So the question is for what kind of data these inequalities hold. Although we still have no geometric characterization of such data, we will show that certain well behaved data, to be explained below, are included. Since $\Delta > 0$, for the above two inequalities to hold it is sufficient that $X_1^T AT_0 \geq 0$ and $-X_0^T AT_1 \geq 0$, which can be rewritten as

$$(X_1 - X_0)^T T_0 \geq 0$$

and

$$(X_1 - X_0)^T T_1 \geq 0,$$

since $X_0^T AT_0 = 0$, $X_1^T AT_1 = 0$ and the last components of $X_1 - X_0$, T_0 and T_1 are zero. The geometric interpretation of these two inequalities is that the angle between the vector $\overrightarrow{X_0X_1}$ and the tangent direction T_0 and the angle between $\overrightarrow{X_0X_1}$ and the tangent direction T_1 are both not greater than $\pi/2$. Clearly, it is reasonable to consider data $D = \{X_0, T_0, X_1, T_1\}$ satisfying this property as well behaved. For regular data that does not satisfy this property, i.e. not well behaved data, it is possible that $k_0^{(2)}$ or $k_1^{(2)}$ is negative. But in this case, since either the angle between $\overrightarrow{X_0X_1}$ and T_0 or the angle between $\overrightarrow{X_0X_1}$ and T_1 is greater than $\pi/2$, the possible use of major arcs in the interpolating biarc is justifiable in some degree.

For singular data a spherical biarc with equal chords is generated if the joint point Z is chosen from $(J - \{X_0, X_1\}) \bigcap L$, where J is the $(d-2)$-dimensional sphere specified in Theorem 5.3.10. Since in this case the two control polygons of the two arcs of the biarc are congruent (because $Y_0 = Y_1$), the two arcs also have equal radii. Note that for singular data D, there exist no proper spherical biarcs interpolating D consisting only of minor arcs.

6.1.3 Minimal twist biarc

In this subsection we will derive a formula to compute a spherical biarc with minimal twist interpolating regular data D on S^{d-1}. The main idea of this derivation is based on an insight in [NuM88]. First we will investigate the spherical biarc with the minimum twist, or twist angle. Let C_0 and C_1 be the two arcs of a spherical biarc B interpolating D. The twist angle of a spherical biarc is defined to be the angle ϕ between the two planes containing arcs C_0 and C_1, where we take the angle ϕ to be the smaller of the two supplementary angles thus formed, i.e. $0 \leq |\phi| \leq \pi/2$. Below we will only discuss the minimum twist biarc solution for regular data. The problem of finding minimum twist spherical biarcs for singular data has not yet been solved.

The explicit solution to the minimum twist angle biarc, to be described below, involves simple formulas that are easy to compute but relatively difficult to derive. Consider regular data $D = \{X_0, T_0, X_1, T_1\}$ on S^{d-1}. Then the two parameters k_0 and k_1 of any spherical biarc for D are related by (5.10). We will show first that $\tan^2\phi$ is a simple function of $k_0 + k_1$; this is the technique used for $3D$ space biarcs in [NuM88]. Then we will translate the minimization problem for ϕ to that of minimizing $|k_0 + k_1|$.

Lemma 6.1.1: *Let $D = \{X_0, T_0, X_1, T_1\}$ be regular data and let ϕ be the twist angle of a proper spherical biarc B interpolating D on S^{d-1} given by k_0 and k_1 satisfying (5.10), where $k_0 k_1 \neq 0$. Then*

$$\tan^2\phi = \frac{a}{[\Delta/(k_0+k_1) + X_1^T A T_0 - X_0^T A T_1]^2},$$

where

$$\begin{aligned} a &= -2(T_0^T A T_1)(X_0^T A T_1)(X_1^T A T_0) - (X_0^T A T_1)^2 \\ &\quad -(X_1^T A T_0)^2 - 2[1 - (T_0^T A T_1)^2](X_0^T A X_1), \end{aligned}$$

and

$$\Delta = (X_0^T A X_1)(T_0^T A T_1 - 1) - (X_0^T A T_1)(X_1^T A T_0).$$

PROOF: First we compute $\cos^2\phi$. Let X_0Y_0Z and ZY_1X_1 be the Bézier control polygons of the two arcs of biarc B, where Y_0 and Y_1 are given by (5.3), and the joint Z is given by (5.9). By Lemma 5.3.2, $Y_0 \neq Y_1$, where Y_0 and Y_1 are in normalized form; therefore the straight line Y_0Y_1 is well defined. Let R_0 and R_1 be points in normalized form on the straight line Y_0Y_1 so that both lines X_0R_0 and X_1R_1 are perpendicular to Y_0Y_1. Denote $V_0 = X_0 - R_0$ and $V_1 = X_1 - R_1$; then V_0 and V_1 are two points at infinity in E^d since X_0, R_0, X_1 and R_1 are in normalized form. So V_0 and V_1 can be regarded as two vectors making the twist angle ϕ, i.e. $\cos\phi = V_0^T V_1/(|V_0|\ |V_1|)$, where for a point V at infinity we define $|V| = \sqrt{V^T V}$.

Since $Y \equiv Y_1 - Y_0 \neq 0$, $W = Y/|Y|$ is well defined. As the projection of $X_0 - Y_0$ on the direction W is $R_0 - Y_0$, we have

$$R_0 - Y_0 = [(X_0 - Y_0)^T W]W.$$

Thus

$$\begin{aligned} V_0 &= X_0 - R_0 = (X_0 - Y_0) - (R_0 - Y_0) \\ &= X_0 - Y_0 - [(X_0 - Y_0)^T W]W. \end{aligned}$$

Similarly,

$$V_1 = X_1 - Y_1 - [(X_1 - Y_1)^T W]W.$$

Since $Y_0 = X_0 + k_0 T_0$ and $Y_1 = X_1 - k_1 T_1$, we obtain

$$V_0 = -k_0 T_0 + k_0 (T_0^T W)W,$$

and

$$|V_0|^2 = V_0^T V_0 = k_0^2 [1 - (T_0^T W)^2].$$

Similarly

$$V_1 = k_1 T_1 - k_1 (T_1^T W)W,$$

and consequently,

$$|V_1|^2 = k_1^2 [1 - (T_1^T W)^2],$$

and

$$V_0^T V_1 = k_0 k_1 [(T_0^T W)(T_1^T W) - (T_0^T T_1)].$$

Then, recalling that $W = Y/|Y|$, where $Y = Y_1 - Y_0$, we have

$$\begin{aligned} \cos^2\phi &= \frac{(V_0^T V_1)^2}{|V_0|^2 |V_1|^2} \\ &= \frac{[(T_0^T W)(T_1^T W) - (T_0^T T_1)]^2}{[1 - (T_0^T W)^2][1 - (T_1^T W)^2]} \\ &= \frac{[(T_0^T Y)(T_1^T Y) - (T_0^T T_1)(Y^T Y)]^2}{[Y^T Y - (T_0^T Y)^2][Y^T Y - (T_1^T Y)^2]}. \end{aligned}$$

So

$$\tan^2\phi = \frac{1 - \cos^2\phi}{\cos^2\phi}$$

$$\begin{aligned}
&= \frac{[Y^TY-(T_0^TY)^2][Y^TY-(T_1^TY)^2]-[(T_0^TY)(T_1^TY)-(T_0^TT_1)(Y^TY)]^2}{[(T_0^TY)(T_1^TY)-(T_0^TT_1)(Y^TY)]^2}\\
&= \frac{[Y^TAY-(T_0^TAY)^2][Y^TAY-(T_1^TAY)^2]-[(T_0^TAY)(T_1^TAY)-(T_0^TAT_1)(Y^TAY)]^2}{[(T_0^TAY)(T_1^TAY)-(T_0^TAT_1)(Y^TAY)]^2}\\
&\equiv \frac{NUM}{DENOM}.
\end{aligned}$$

Now we will simplify this expression to express $\tan^2\phi$ as a function of the sum k_0+k_1. By (5.3) and (5.6), which is equivalent to (5.10),

$$\begin{aligned}
Y^TAY &= (Y_1-Y_0)^TA(Y_1-Y_0)\\
&= Y_1^TAY_1+Y_0^TAY_0-2Y_0^TAY_1\\
&= k_0^2+k_1^2+2k_0k_1\\
&= (k_0+k_1)^2.
\end{aligned}$$

Substituting in $Y=Y_1-Y_0=X_1-X_0-k_0T_0-k_1T_1$ and regrouping, we obtain

$$\begin{aligned}
NUM &= (k_0+k_1)^2\Big\{-2(T_0^TAT_1)(X_0^TAT_0)(X_1^TAT_0)-(X_0^TAT_1)^2-(X_1^TAT_0)^2\\
&\quad -2[1-(T_0^TAT_1)^2](X_0^TAX_1)\\
&\quad +2[1-(T_0^TAT_1)^2][X_0^TAX_1+k_0X_1^TAT_0-k_1X_0^TAT_1+k_0k_1(1-T_0^TAT_1)]\Big\}\\
&= (k_0+k_1)^2a,
\end{aligned}$$

where

$$\begin{aligned}
a &= -2(T_0^TAT_1)(X_0^TAT_1)(X_1^TAT_0)-(X_0^TAT_1)^2-(X_1^TAT_0)^2\\
&\quad -2[1-(T_0^TAT_1)^2](X_0^TAX_1).
\end{aligned}$$

Note that the last term drops out due to (5.10).

Now consider the denominator $DENOM$. Substituting in $Y=X_1-X_0-k_0T_0-k_1T_1$ and simplifying yield

$$\begin{aligned}
\sqrt{DENOM} &= -(X_0^TAT_1)(X_1^TAT_0)+k_0[T_1^TAX_0-(T_0^TAT_1)(T_0^TAX_1)]\\
&\quad +k_1[-T_0^TAX_1+(T_0^TAT_1)(T_1^TAX_0)]+k_0k_1[1-(T_0^TAT_1)]^2.
\end{aligned}$$

Using (5.10), we substitute

$$k_0k_1(1-T_0^TAT_1)=-X_0^TAX_1-k_0X_1^TAT_0+k_1X_0^TAT_1,$$

for the last term in the above expression and obtain

$$\begin{aligned}
\sqrt{DENOM} &= -(X_0^TAT_1)(X_1^TAT_0)-(1-T_0^TAT_1)(X_0^TAX_1)\\
&\quad +(k_0+k_1)(X_0^TAT_1-X_1^TAT_0)\\
&= \Delta+(k_0+k_1)(X_0^TAT_1-X_1^TAT_0),
\end{aligned}$$

where

$$\Delta = (X_0^T A X_1)(T_0^T A T_1 - 1) - (X_0^T A T_1)(X_1^T A T_0).$$

So

$$\tan^2\phi = \frac{NUM}{DENOM} = \frac{a}{[\Delta/(k_0+k_1) + X_0^T A T_1 - X_1^T A T_0]^2}.$$

□

Since $|\phi|$ is not greater than $\pi/2$, to minimize ϕ we just need to minimize $\tan^2\phi$ over $k_0 + k_1$. That is, to maximize $|\Delta/(k_0+k_1) + p|$, where $p = X_0^T A T_1 - X_1^T A T_0$. Now we show that for this maximum to be attained it is necessary (but not sufficient) that a local minimum of $|k_0+k_1|$ is attained. First, the maximum $|\Delta/(k_0+k_1)+p|$ is a local maximum. When $\Delta/(k_0+k_1)$ and p are of the same sign or $\Delta/(k_0+k_1)$ and p are of opposite signs (including $p = 0$) and $|\Delta/(k_0+k_1)| \geq |p|$, obviously, $|k_0+k_1|$ is a local minimum if and only if $|\Delta/(k_0+k_1)+p|$ is a local maximum; if $\Delta/(k_0+k_1)$ and p are of the opposite signs and $|\Delta/(k_0+k_1)| < |p|$ then $|\Delta/(k_0+k_1)+p|$ can not be the maximum for simply choosing another pair of k_0' and k_1' such that $(k_0+k_1)(k_0'+k_1') < 0$, which is always possible as can be seen in Fig. 5.4.1, we would have

$$|\Delta/(k_0+k_1)+p| < |p| < |\Delta/(k_0'+k_1')+p)|,$$

which is a contradiction.

In the proof of Theorem 5.4.4 it is shown that the straight line $k_0 + k_1 = 0$ in the k_0-k_1 plane does not intersect the hyperbola (5.10) (see Fig. 5.4.1). That is, the straight line $k_0 + k_1 = 0$ lies between the two branches of the hyperbola. Therefore $k_0 + k_1 \neq 0$ and the two local minima of $|k_0 + k_1|$ are attained on the two branches of the hyperbola (5.10). It is clear that $|k_0 + k_1|$ has no maximum over the hyperbola.

Let $S(k_0) = k_0 + k_1$. Solving (5.10) for k_1 in terms of k_0 yields

$$S(k_0) = k_0 + \frac{X_0^T A X_1 + k_0 X_1^T A T_0}{X_0^T A T_1 + k_0(T_0^T A T_1 - 1)}. \tag{6.2}$$

The two solutions of $dS(k_0)/dk_0 = 0$ are easily found to be

$$k_0^{(1)} = \frac{-X_0^T A T_1 + \sqrt{\Delta}}{T_0^T A T_1 - 1} \tag{6.3}$$

and

$$k_0^{(2)} = \frac{-X_0^T A T_1 - \sqrt{\Delta}}{T_0^T A T_1 - 1}, \tag{6.4}$$

and the corresponding k_1, solved for from (5.10), are

$$k_1^{(1)} = \frac{X_1^T A T_0 + \sqrt{\Delta}}{T_0^T A T_1 - 1} \tag{6.5}$$

and

$$k_1^{(2)} = \frac{X_1^T A T_0 - \sqrt{\Delta}}{T_0^T A T_1 - 1}. \tag{6.6}$$

The fact that $\Delta > 0$ is indicated by the remark following the proof of Theorem 5.4.4.

The following theorem summarizes the procedure for determining the minimum twist angle solution k_0, k_1.

Theorem 6.1.2: *Using the above notation, if*

$$\left|\frac{\Delta}{S(k_0^{(1)})} + p\right| \geq \left|\frac{\Delta}{S(k_0^{(2)})} + p\right|,$$

then the solution (k_0, k_1) that minimizes the twist angle is $(k_0^{(1)}, k_1^{(1)})$; otherwise $(k_0, k_1) = (k_0^{(2)}, k_1^{(2)})$.

The hyperbola determined by $F(k_0, k_1) = 0$ (see Eqn. (5.15)) is a translation of a hyperbola $k_0 k_1 = r > 0$. As indicated in the proof of Theorem 5.4.4, this hyperbola does not intersect the straight line $k_0 + k_1 = 0$. On the other hand, we know that the two local minimal twist solutions $(k_0^{(1)}, k_1^{(1)})$ and $(k_0^{(2)}, k_1^{(2)})$ are two points on the two branches of the hyperbola. Since in (6.3) and (6.5) the denominator $T_0^T A T_1 - 1 < 0$ for regular data, it follows that $k_0^{(1)} < k_0^{(2)}$ and $k_1^{(1)} < k_1^{(2)}$, that is $(k_0^{(1)}, k_1^{(1)})$ must a point on the left branch of the hyperbola. Thus at least one of $k_0^{(1)}$ and $k_1^{(1)}$ is negative. Since a negative k_0 or k_1 gives a major arc in a biarc, which is undesirable for its relatively large winding, in applications we may want to consider only the other solution $(k_0^{(2)}, k_1^{(2)})$ given by (6.4) and (6.6), which has more chance to be a biarc consisting of two minor arcs. Experiments showed that this is not always the case, i.e. for some regular data $(k_0^{(2)}, k_1^{(2)})$ is not in the first quadrant of the k_0-k_1 plane. The situation is illustrated in Fig. 5.4.1, i.e. no point of the hyperbola $F(k_0, k_1) = 0$ in the first quadrant attains a local maximum distance to the straight line $k_0 + k_1 = 0$. So if only minor arcs can be used for constructing biarcs, the minimum twist biarc problem may not have a solution. In the following discussion by the minimal twist biarc we will mean the biarc with the joint determined by $k_0^{(2)}$ and $k_1^{(2)}$ given by (6.4) and (6.6).

In the next section we will see that the minimal twist biarc tends to give a large difference between the two radii and a large winding compared with the equal chord biarc.

6.2 Evaluation of three choices

In Section 6.1 two simple formulae have been derived for computing the joint of a spherical biarc interpolating regular data D on S^{d-1}. Together with the solution $k_0 = k_1 > 0$ given in (5.16), there are now the following three choices of joints available:

(1) *Choice* 1: The equal chord biarc:

$$\begin{aligned} k_0 &= \frac{-X_0^T A X_1}{X_1^T A T_0 + \sqrt{\Delta}}, \\ k_1 &= \frac{-X_0^T A X_1}{-X_0^T A T_1 + \sqrt{\Delta}}, \end{aligned} \tag{6.7}$$

where

$$\Delta = (X_0^T A X_1)(T_0^T A T_1 - 1) - (X_0^T A T_1)(X_1^T A T_0).$$

(2) *Choice* 2: The equal k biarc:

$$k_0 = k_1 = \frac{-b + \sqrt{b^2 - 4ac}}{2a}, \tag{6.8}$$

where

$$a = 1 - T_0^T A T_1, \quad b = X_1^T A T_0 - X_0^T A T_1, \quad c = X_0^T A X_1.$$

(3) *Choice* 3: The minimal twist biarc:

$$\begin{aligned} k_0 &= \frac{X_0^T A T_1 + \sqrt{\Delta}}{1 - T_0^T A T_1}, \\ k_1 &= \frac{-X_1^T A T_0 + \sqrt{\Delta}}{1 - T_0^T A T_1}, \end{aligned} \tag{6.9}$$

where Δ is defined as above.

In this section these three choices for k_0, k_1 are used to compute the joint of a biarc to interpolate randomly generated data on S^2, and the ratio of the radii of each biarc is computed. Each choice will then be judged by the statistics thus obtained. Our experimental results suggest that the equal chord biarc is the best choice of the three. In the following the design of the experiment is explained; then the experimental results are presented and analyzed.

Let r_0 and r_1 be the radii of the first and second arcs of a spherical biarc interpolating $D = \{X_0, T_0, X_1, T_1\}$, i.e. the arc starting at X_0 and the arc ending at X_1. The random variable that is observed in the experiment is $R = \min\{r_0/r_1,\ r_1/r_0\}$, instead of the ratio

r_0/r_1. R will be called the *radius ratio* of the biarc. Note that the range of the radius ratio R is $[0,\ 1]$, and R is an upper bound for the smaller radius since the largest circle on S^2 has radius 1. When R is close to 1, the two radii are close and a spherical biarc with fair shape is expected.

The following explains why the ratio r_0/r_1 is not observed. Suppose for a biarc we have $r_0/r_1 = 100$ and for another biarc we have $r_0/r_1 = 0.01$. These two ratios represent the same fact as far as we are concerned; that is, the radius of one arc is 100 times greater than that of the other. But when we compute the mean or the standard deviation of r_0/r_1, the above two values obviously yield different contributions, which is an undesirable result.

Three experiments have been performed to compare the above three choices for the joint. In each experiment up to 50000 sets of data $D = \{X_0, T_0, X_1, T_1\}$ are randomly generated on $S^2 \subset E^3$. For each of these 50000 data sets the three formulae listed at the beginning of this section are used to calculate three joints of biarcs. Then the radius ratio R is computed for each of the three biarcs, with the three radius ratios being denoted by R_1, R_2 and R_3. Our intention is to find out in the worst case how small the radius ratio resulting from each formula can be, and on average which formula gives the largest radius ratio. We are also interested in the ratio of the arclength of a biarc to the distance $|X_0X_1|$ and the ratio of the smaller radius of a biarc to $|X_0X_1|$, because the former measures the winding of the biarc and the latter tells whether or not the biarc has a sharp turn when the distance $|X_0X_1|$ is small. In the three experiments the data $D = \{X_0, T_0, X_1, T_1\}$ are such that $|X_0X_1|$ is not restricted, $|X_0X_1| > 0.4$, and $|X_0X_1| < 0.1$, respectively.

The purposes of these three experiments are as follows. In the first experiment general observations on the distributions of radius ratios R_1, R_2 and R_3 were made. As indicated in the last section a very small radius ratio alone does not in general imply that the biarc has a sharp turn. So in the second experiment random data were generated such that $|X_0X_1| > 0.4$, in which case we can tell that a biarc has a sharp turn if its radius ratio is very small. In order to investigate the case where the two points X_0 and X_1 are very close on S^2, in the third experiment the data were generated such that $|X_0X_1| < 0.1$. In this case, there is a sharp turn on a biarc if the ratio of the smaller radius of the biarc to the distance $|X_0X_1|$ is very small, and a biarc has an excessive winding if the ratio of the arclength of the biarc to $|X_0X_1|$ is very large.

The fact that the experiments were carried out only on S^2 does not affect the generality of the results, since given noncoplanar regular data D on $S^{d-1} \subset E^d$, all the spherical biarcs interpolating D are contained in the 3-dimensional affine manifold determined by D, therefore contained in a 2-dimensional sphere on S^{d-1}.

Random data $D = \{X_0, T_0, X_1, T_1\}$ are generated as follows. In the first experiment, to obtain a random point on S^2, first a random point that is uniformly distributed in the cube $[-10, 10]^3$ is generated, and then this point is projected onto the sphere S^2 from the center of the sphere. To generate a random tangent direction $T = [t_0, t_1, t_2, 0]^T$ at a point $U \in S^2$, t_0 and t_1 are generated as uniformly distributed numbers in $[-10, 10]$, and t_2 is determined such that $U^T AT = 0$ is satisfied, where $A = \text{diag}[1,\ 1,\ 1,\ -1]$. Then T is normalized so that $T^T AT = 1$. In this way X_0 and X_1 and T_0 and T_1 are generated. In the second experiment once X_0 is obtained, X_1 is randomly generated subject to the condition that $|X_0X_1| > 0.4$. In the third experiment the data are generated similarly.

In each experiment the means and standard deviations of the three radius ratios R_1, R_2 and R_3 were recorded, denoted by $mean\_R_1$, $mean\_R_2$, $mean\_R_3$, $stdv\_R_1$, $stdv\_R_2$, and $stdv\_R_3$, respectively. These data are shown in Table 6.2.1, 6.2.4, and 6.2.7 for the three experiments. Each row is the observed means and standard deviations for the sample of indicated size, from 100 to 50000. Also recorded are the minima of the radius ratios over the sample, denoted by min_R_1, min_R_2, and min_R_3, which are shown in Table 6.2.2, 6.2.5, and 6.2.8 for the three experiments. Finally the maxima of the ratio of the arclength of a biarc to the distance $|X_0X_1|$, denoted by max_ad_i, $i = 1, 2, 3$, and the minima of the ratio of the smaller radius of a biarc to $|X_0X_1|$, denoted by min_rd_i, $i = 1, 2, 3$, are listed in Table 6.2.3, 6.2.6, and 6.2.9 for the three experiments. In all these tables, a sample of certain size is a subset of a sample of a larger size.

In the first experiment, in Table 6.2.1, note that $mean\_R_1$ is significantly greater than $mean\_R_2$ and $mean\_R_3$; also the standard deviations indicate that R_1 has less fluctuation than R_2 and R_3. The latter may be in part due to the fact that $mean\_R_1$ is closer to the boundary of the range $[0,\ 1]$ than $mean\_R_2$ and $mean\_R_3$. Table 6.2.2 shows that the radius ratios R_2 and R_3 are more inclined to give extremely small values than R_1. For example, for the sample of 50000 data, $min\_R_1 = 0.023410$, $min\_R_2 \approx 0$, and $min\_R_3 = 0.000002$. Finally, Table 6.2.3 indicates that, as expected, on average the second choice gives the smallest arclength to distance ratio, i.e. the smallest winding in the worst case among the three biarcs, and the equal chord biarc gives the second smallest winding, while the minimal twist biarc gives the largest winding. From the min_rd_i, $i = 1, 2, 3$, we see that unlike the equal chord biarcs the other two biarcs do have very sharp turns in the worst case.

In the second experiment, the data $D = \{X_0, T_0, X_1, T_1\}$ are such that $|X_0X_1| > 0.4$. From Table 6.2.4, on average the first choice still gives the largest radius ratio. From Table 6.2.5, for the sample of 50000 data, $min\_R_2 = 0.000004$ and $min\_R_3 = 0.000020$, so we can conclude that in the worst case the second and third choices can give a biarc with an extremely sharp turn on it even when $|X_0X_1|$ is relatively large; this is also confirmed by

size of sample	$mean\_R_1$	$mean\_R_2$	$mean\_R_3$	$stdv\_R_1$	$stdv\_R_2$	$stdv\_R_3$
100	0.797128	0.566336	0.566900	0.186403	0.311543	0.304779
200	0.786460	0.599301	0.581998	0.194265	0.299317	0.310807
500	0.791916	0.580340	0.581721	0.187589	0.304658	0.301960
1000	0.790914	0.573618	0.576835	0.194869	0.307181	0.304135
2000	0.798489	0.570528	0.576780	0.188689	0.305348	0.299257
5000	0.797896	0.565220	0.573587	0.190352	0.305769	0.300509
10000	0.796198	0.564239	0.568783	0.191140	0.304847	0.301012
20000	0.797505	0.561132	0.569268	0.191134	0.306732	0.301589
50000	0.799047	0.564019	0.571974	0.189523	0.307232	0.302413

Table 6.2.1 $mean\_R_i$ **and** $stdv\_R_i$: The means and standard deviations of the three radius ratios R_i, $i = 1, 2, 3$, for random data on S^2.

size of sample	min_R_1	min_R_2	min_R_3
100	0.187908	0.000096	0.007883
200	0.187908	0.000096	0.006189
500	0.172216	0.000096	0.003759
1000	0.115345	0.000096	0.000941
2000	0.115345	0.000096	0.000014
5000	0.078263	0.000096	0.000014
10000	0.070873	0.000096	0.000014
20000	0.023410	0.000038	0.000014
50000	0.023410	0.000000	0.000002

Table 6.2.2 min_R_i: The minima of the three radius ratios R_i, $i = 1, 2, 3$, for random data on S^2.

size of sample	max_ad_1	max_ad_2	max_ad_3	min_rd_1	min_rd_2	min_rd_3
100	8.383589	4.458478	6.520546	0.273540	0.000298	0.004565
200	24.825400	24.776892	24.776580	0.253849	0.000298	0.003837
500	24.825400	24.776892	24.776580	0.253849	0.000298	0.002059
1000	24.825400	24.776892	77.002241	0.253849	0.000298	0.000523
2000	29.425547	29.361194	77.002241	0.251326	0.000298	0.000008
5000	30.812709	29.361194	77.002241	0.251080	0.000236	0.000008
10000	86.054463	53.528423	88.426479	0.250883	0.000076	0.000008
20000	86.054463	53.528423	104.138630	0.250883	0.000022	0.000008
50000	119.495479	96.685922	174.438704	0.250883	0.000000	0.000001

Table 6.2.3 max_ad_i **and** min_rd_i: The maxima of arc-distance ratios ad_i and the minima of radius-distance ratios rd_i of the three biarcs for random data on S^2, $i = 1, 2, 3$.

the min_rd_i's in Table 6.2.6. Note that since the largest circle on S^2 has radius one, the radius ratio actually gives an upper bound on the smaller radius of the biarc. So when $R_2 = 0.000004$ and $R_3 = 0.000020$, the two corresponding arcs with the radii ≤ 0.000004 and ≤ 0.000020, respectively, are too small relative to the distance $|X_0X_1| > 0.4$.

In the same setting we have $min\_R_1 = 0.137636$ for the equal chord biarc, which is considerably large compared with min_R_2 and min_R_3. In fact, we can show that with $|X_0X_1| > 0.4$ no very small radius ratio can be generated by the equal chord biarc. For, let $l = |X_0X_1| \leq 1$ and $d = |MX_0|$ where M is a point on S^2 such that $|MX_0| = |MX_1|$ and $|MX_0|$ is the minimum. Then, referring to Fig. 6.2.1, it is straightforward to obtain

$$\begin{aligned} d &= 2\sin(\theta/2) = 2\sqrt{\frac{1-\cos\theta}{2}} \\ &= 2\sqrt{\frac{1-\sqrt{1-\sin^2\theta}}{2}} = 2\sqrt{\frac{1-\sqrt{1-(l/2)^2}}{2}}. \end{aligned}$$

Since $l > 0.4$, $d > 0.201018$. Let the two radii of a biarc interpolating D be r_0 and r_1 and, without loss of generality, assume $r_0 \leq r_1$. Let Z, with $|ZX_0| = |ZX_1|$, be the joint given by choice 1. Using the triangle inequality it is clear that $2r_0 \geq |ZX_0| \geq |MX_0| = d$, and $r_1 \leq 1$. Then the radius ratio

$$R_1 = \frac{r_0}{r_1} \geq \frac{d}{2} > 0.100509.$$

Therefore $min\_R_1 > 0.100509$, which conforms with the min_R_1 in Table 6.2.5.

Finally, the max_ad_i, $i = 1, 2, 3$, in Table 6.2.6 support the same conclusion derived

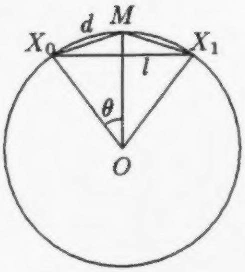


Figure 6.2.1.

from Table 6.2.3. And the min_rd_i, $i = 1, 2, 3$, confirm again the fact that very sharp turns can be found in biarcs given by the choices 2 and 3.

The results of the third experiment are presented in Table 6.2.7, 6.2.8, and 6.2.9. First of all, Table 6.2.7 shows that in this case, i.e. when the X_0 and X_1 are very close to each other, there is not as much difference among $mean\_R_i$, $i = 1, 2, 3$, as shown in the first two experiments. Therefore we conclude that only when the distance $|X_0X_1|$ is not very small, the equal chord biarc tends to give a greater radius ratio than the other two biarcs, while when $|X_0X_1|$ is very small, $mean\_R_1$ is about the same as $mean\_R_3$, and $mean\_R_2$ is the smallest. In Table 6.2.8, we see that extremely small radius ratios can be observed in all the three choices of the joints, but since $|X_0X_1|$ is small in this case, this does not indicate that the biarc with the shown small radius ratio necessarily has a sharp turn. Table 6.2.9 contains the most significant information about the shape of the biarc. From the max_ad_i, $i = 1, 2, 3$, we see that the second choice gives smallest windings in the worst case as expected; actually, the same ordering of performances of the three choices with respect to the winding are kept here as in the first two experiments. Finally, from the min_rd_i, $i = 1, 2, 3$ in Table 6.2.9, note that even relative to the distance $|X_0X_1|$, choices 2 and 3 still give biarcs with an arc of extremely small radius. But at the same time it seems that min_rd_1 is bounded below by 1/4. In fact, it is easy to see that this is indeed the case, since $4r > l$ follows from the triangle inequality, where r is the smaller radius of the biarc and $l = |X_0X_1|$. Hence, relative to $|X_0X_1|$ the two radii of an equal chord biarc can not be very small.

Based on the above experiments and analysis we conclude that, if the sharp turning of a biarc is the major concern in the shape control, the equal chord biarc gives better results

size of sample	$mean\_R_1$	$mean\_R_2$	$mean\_R_3$	$stdv\_R_1$	$stdv\_R_2$	$stdv\_R_3$
100	0.766125	0.546021	0.530648	0.210087	0.301873	0.320598
200	0.792935	0.557503	0.553674	0.191862	0.305203	0.321075
500	0.798240	0.577798	0.567308	0.181221	0.303723	0.318409
1000	0.803513	0.581572	0.566898	0.181370	0.303017	0.314929
2000	0.810874	0.577990	0.574894	0.174460	0.303608	0.305918
5000	0.807002	0.573175	0.573695	0.178951	0.305068	0.303168
10000	0.807611	0.572483	0.574661	0.179088	0.306292	0.302371
20000	0.806647	0.569364	0.571955	0.179189	0.304894	0.302311
50000	0.807685	0.569445	0.571950	0.179042	0.305278	0.303046

Table 6.2.4 $mean\_R_i$ **and** $stdv\_R_i$ **with** $|X_0X_1| > 0.4$: The means and standard deviations of the three radius ratios R_i, $i = 1, 2, 3$, for random data on S^2 for which $|X_0X_1| > 0.4$.

size of sample	min_R_1	min_R_2	min_R_3
100	0.254670	0.003057	0.000594
200	0.167321	0.001391	0.000594
500	0.167321	0.001391	0.000594
1000	0.167321	0.001391	0.000594
2000	0.153422	0.001391	0.000594
5000	0.137636	0.000461	0.000227
10000	0.137636	0.000086	0.000227
20000	0.137636	0.000054	0.000026
50000	0.137636	0.000004	0.000020

Table 6.2.5 min_R_i **with** $|X_0X_1| > 0.4$: The minima of the three radius ratios R_i, $i = 1, 2, 3$, for random data on S^2 for which $|X_0X_1| > 0.4$.

size of sample	max_ad_1	max_ad_2	max_ad_3	min_rd_1	min_rd_2	min_rd_3
100	8.383589	7.389397	12.435508	0.253684	0.002126	0.000353
200	8.383589	7.389397	12.435508	0.253684	0.001118	0.000353
500	9.703329	8.486590	12.435508	0.253458	0.001118	0.000353
1000	13.491321	10.969571	12.435508	0.253458	0.001118	0.000353
2000	13.491321	11.290834	14.284111	0.253166	0.001027	0.000353
5000	13.644763	12.906835	14.284111	0.252841	0.000236	0.000127
10000	13.983047	13.177054	14.732622	0.251967	0.000045	0.000127
20000	14.727013	14.711459	14.732622	0.251967	0.000031	0.000014
50000	15.504469	14.711459	14.732622	0.251946	0.000002	0.000014

Table 6.2.6 max_ad_i **and** min_rd_i **with** $|X_0X_1| > 0.4$: The maxima of arc-distance ratios ad_i and the minima of radius-distance ratios rd_i of the three biarcs for random data on S^2 for which $|X_0X_1| > 0.4$, $i = 1, 2, 3$.

size of sample	$mean\_R_1$	$mean\_R_2$	$mean\_R_3$	$stdv\_R_1$	$stdv\_R_2$	$stdv\_R_3$
100	0.513139	0.339255	0.508100	0.286391	0.299262	0.281459
200	0.518718	0.360062	0.514996	0.292357	0.306163	0.290297
500	0.535859	0.382144	0.534002	0.298900	0.313135	0.298035
1000	0.544639	0.395551	0.542258	0.299421	0.315552	0.299293
2000	0.558113	0.405809	0.555884	0.297544	0.318138	0.297370
5000	0.555589	0.400917	0.553016	0.299202	0.316001	0.299083
10000	0.555950	0.398597	0.553320	0.299415	0.314910	0.299447
20000	0.562228	0.400922	0.559324	0.299370	0.316738	0.299587
50000	0.561702	0.401452	0.558801	0.299364	0.316189	0.299633

Table 6.2.7 $mean\_R_i$ **and** $stdv\_R_i$ **with** $|X_0X_1| < 0.1$: The means and standard deviations of the three radius ratios R_i, $i = 1, 2, 3$, for random data on S^2 for which $|X_0X_1| < 0.1$.

size of sample	min_R_1	min_R_2	min_R_3
100	0.024499	0.001649	0.024449
200	0.024499	0.000395	0.024449
500	0.007879	0.000036	0.004666
1000	0.005920	0.000036	0.001353
2000	0.003982	0.000036	0.001003
5000	0.002822	0.000036	0.000240
10000	0.002822	0.000036	0.000138
20000	0.001051	0.000036	0.000124
50000	0.001051	0.000008	0.000040

Table 6.2.8 min_R_i **with** $|X_0X_1| < 0.1$: The minima of the three radius ratios R_i, $i = 1, 2, 3$, for random data on S^2 for which $|X_0X_1| < 0.1$.

size of sample	max_ad_1	max_ad_2	max_ad_3	min_rd_1	min_rd_2	min_rd_3
100	260.269499	24.524743	644.807746	0.261847	0.009274	0.014089
200	795.462196	73.246363	644.807746	0.255050	0.001110	0.014089
500	795.462196	256.545119	6312.642710	0.251316	0.000640	0.003689
1000	795.462196	544.065434	6312.642710	0.250673	0.000640	0.001055
2000	795.462196	544.065434	6312.642710	0.250052	0.000122	0.000512
5000	1448.419133	671.653572	6312.642710	0.250018	0.000122	0.000143
10000	2213.575014	1003.793661	6312.642710	0.250018	0.000101	0.000074
20000	2213.575014	1003.793661	6312.642710	0.250008	0.000101	0.000074
50000	2213.575014	1511.330495	8210.653188	0.250002	0.000010	0.000025

Table 6.2.9 max_ad_i **and** min_rd_i **with** $|X_0X_1| < 0.1$: The maxima of arc-distance ratios ad_i and the minima of radius-distance ratios rd_i of the three biarcs for random data on S^2 for which $|X_0X_1| < 0.1$, $i = 1, 2, 3$.

than the second and third choices, in the sense that it never gives an extremely sharp turn relative to $|X_0X_1|$ and on average it gives the larger radius ratio than the other two choices when $|X_0X_1|$ is not very small. One caution about using the equal chord biarc is that when the interpolated data is not well behaved (cf. the next to last paragraph in Subsection 6.1.2), the biarc may have large winding.

The main reason that the second choice of the joint is not very ideal is that the condition $k_0 = k_1 > 0$ does not have any direct relationship with the radii of the two arcs of the biarc, except that it ensures that the two arcs are minor. As for the third choice, which also tends to give a very sharp turn on a biarc, it should be realized that the original goal of this choice is to give a minimal twist biarc. Therefore, we can only conclude that minimizing the twist and minimizing the difference of the two radii of a spherical biarc are two conflicting goals for some regular data. In an application where the minimal twist is important, the third choice certainly gives the only acceptable biarc.

Fig. 6.2.2 illustrates the biarcs given by the three choices of the joint discussed above interpolating different data. In (a) and (b), the three biarcs all are satisfactory interpolants; they have the same shape in (a) but different shapes in (b). In (c) the minimal twist biarc has a large winding, while in (d) the equal chord biarc has a large winding. In (e) the minimal twist biarc has a sharp turn, and in (f) the equal k biarc has a sharp turn.

6.3 Interpolation of a point sequence

In this section an algorithm is presented for interpolating a point sequence on S^{d-1} using spherical biarcs. The idea of the algorithm is to determine the tangents at the interpolated points so that the shape of the data is preserved. The shape of the data can be regarded as the shape of the spherical polygon obtained by connecting consecutive data points X_i and X_{i+1} by the geodesic line on the sphere which is the minor arc on the great circle on S^{d-1} that passes through X_i and X_{i+1}. Here we assume that X_i and X_{i+1} are not antipodal for otherwise the geodesic line is not uniquely defined.

Algorithm 6.3.1: The input is a point sequence $\{X_i\}_{i=1}^{n}$ in normalized form on S^{d-1}, $d \geq 3$ and $n \geq 3$. Assume that no two consecutive points are antipodal. Let C_i be the minor arc connecting X_i and X_{i+1} on the great circle of S^{d-1} passing through X_i and X_{i+1}, with its direction being from X_i to X_{i+1}. Let T_i^1 be the tangent direction of C_{i-1} at X_i. Let T_i^2 be the tangent direction of C_i at X_i. Let $d_j = \arccos(X_j^T X_{j+1} - 1)$, i.e. the spherical distance between X_j and X_{j+1}, $j = 1, 2, \ldots, n-1$. Note that the X_i's are homogeneous coordinates in normalized form.

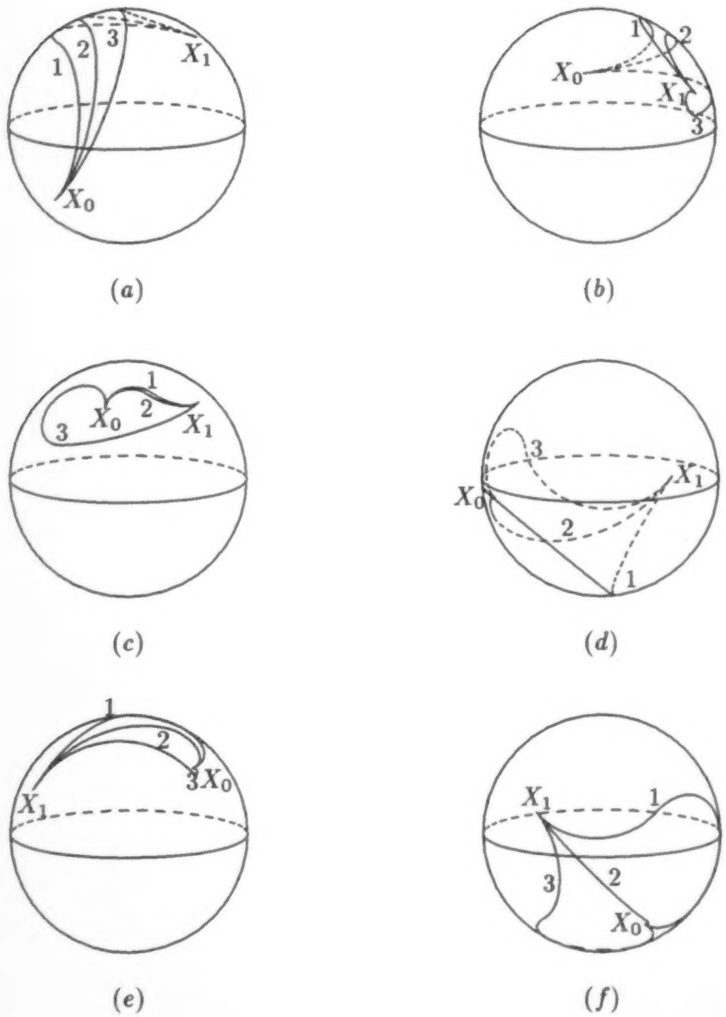
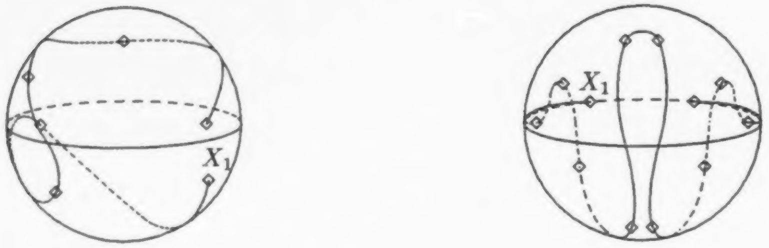


Figure 6.2.2 Interpolations by the three biarcs: Different data configurations interpolated by the three spherical biarcs. The biarcs labelled 1, 2, and 3 are the equal chord biarc, the biarc with $k_0 = k_1 > 0$, and the minimal twist biarc, respectively.



(a) (b)

Figure 6.3.1 Interpolation of data on S^2 by Algorithm 6.3.1.

1. The tangent direction T_i at X_i is defined to be

$$\begin{aligned} T_1 &\leftarrow T_1^2; \\ T_i &\leftarrow \frac{d_{i-1}T_i^1 + d_iT_i^2}{\|d_{i-1}T_i^1 + d_iT_i^2\|}, \quad i = 2, \ldots, n-1; \\ T_n &\leftarrow T_n^1. \end{aligned}$$

2. Apply the spherical biarc to interpolate $D_i = \{X_i, T_i, X_{i+1}, T_{i+1}\}$, $i = 1, \ldots, n-1$. When D_i is regular, use the equal chord biarc where the parameters k_0 and k_1 are given by (6.7). When D_i is singular, use an equal chord biarc with the joint Z being any point in $(J - \{X_i, X_{i+1}\}) \bigcap L$, where J is as defined in Theorem 5.3.10 and $L = \{X | (X_i - X_{i+1})^T AX = 0\}$ is the perpendicular bisecting hyperplane of X_0 and X_1 in E^d (cf. Eqn. (6.1)).□

Fig. 6.3.1 illustrates two examples of interpolating data points on S^2 using Algorithm 6.3.1. In Fig. 6.3.1 (a) the distances between consecutive data points are quite uniform and they are not uniform in (b).

In Algorithm 6.3.1 the equal chord biarcs are used as a result of the study in Section 6.2. As an application requires, biarcs determined by other criteria can also used in Algorithm 6.3.1. We feel that the determination of appropriate tangent directions at the interpolated data points is more important and more crucial to the global quality of the resulting spline than the choice of joints that affects the local shape. Further research is needed in this direction. For instance, how should the tangent directions at the data points and the joints be chosen so that the resulting spline satisfies a certain global optimal goal?

6.4 Interpolation of unit quaternions for orientation modeling

In this section we discuss the application of spherical biarcs for interpolation in the unit quaternion space. This problem has applications in computer animation modeling object orientation and other aspects of computer graphics [Ple89]. We will use spherical biarcs to interpolate a sequence of unit quaternions, and use the difference method discussed in Chapter 2 to evaluate an equidistant sequence of in-between quaternions on the spherical biarc spline curve. As a result, only 4 multiplications and 8 additions/subtractions are needed to compute an in-between quaternion on the interpolating curve, which is a significant speedup over other existing schemes. For instance, in [Sho87] the *Squad* curve is proposed, for which 31 multiplications and several table look-ups are necessary to compute an in-between quaternion.

6.4.1 About quaternions

The use of quaternions for representing object orientations has been known since the relation between quaternions and rotations in E^3 was discovered by Cayley (see [Ple89]). The application of this idea in computer graphics has been explored since 1985 [Sho85, Sho87, Ple89]. Quaternions can be represented as

$$q = w + xi + yj + zk,$$

where w, x, y, z are real numbers, and i, j, k are the imaginary units for the quaternions. Quaternions form a noncommutative ring with addition like that for complex numbers, and multiplication governed by

$$\begin{aligned} i^2 &= j^2 = k^2 = -1, \\ jk &= -kj = i, \\ ki &= -ik = j, \\ ij &= -ji = k. \end{aligned}$$

The norm of a quaternion is defined by

$$|q| = (w^2 + x^2 + y^2 + z^2)^{1/2},$$

and the normalization $q/|q|$, when $q \neq 0$, gives a *unit quaterion*, i.e. a quaternion with unit norm. Unit quaternions can be identified with unit vectors in E^4, or points on the unit sphere $S^3 \subset E^4$.

Like 3×3 orthogonal matrices M with $\det(M) = 1$, unit quaternions can be used to represent rotations in E^3. A rotation around the unit direction vector $[u, v, w]^T$ through the angle θ can be represented by the unit quaternion

$$q = \cos(\theta/2) + \sin(\theta/2)(ui + vj + wk).$$

And conversely, any unit quaternion can be written uniquely in the above form with $-\pi < \theta \leq \pi$, and this represents the corresponding rotation. Since

$$-q = \cos[(2\pi - \theta)/2] + \sin[(2\pi - \theta)/2](-ui - vj - wk),$$

and the rotation around the axis $[-u, -v, -w]^T$ through the angle $2\pi - \theta$ is the same as the rotation around the axis $[u, v, w]$ through the angle θ, the two antipodal unit quaternions $-q$ and q represent the same rotation.

An important fact about the above correspondence is that the product of two unit quaternions gives the same rotation as the composition of the two rotations corresponding to the two original quaternions. This property enables one to use quaternions to formulate and solve problems involving rotations in E^3. Since a unit quaternion contains only four dependent parameters, it provides a more compact tool to describe a rotation than the orthogonal matrix. The following fact is also responsible for the application of unit quaternions in modeling object orientation: two orientations of an object in E^3 differ merely by a rotation in E^3, ignoring any translation. Therefore unit quaternions can be used to represent orientations of an object in E^3 with respect to a reference orientation of the object.

Given a unit quaternion $q = w + xi + yj + zk$, the corresponding orthogonal matrix is

$$\begin{bmatrix} w^2 + x^2 - y^2 - z^2 & 2xy - 2wx & 2xz + 2wy \\ 2xy + 2wy & w^2 - x^2 + y^2 - z^2 & 2yz - 2wx \\ 2xy - 2wy & 2yz + 2wx & w^2 - x^2 - y^2 + z^2 \end{bmatrix}.$$

Conversely, using combinations of entries in a rotation matrix, one can extract the corresponding quaternion. A procedure that avoids numerical exceptions for the conversion from a rotation matrix to a pair of antipodal quaternions is given in [Sho85]. For more properties of quaternions, the reader is referred to [Alt86, Sho85, Ple89].

6.4.2 Orientation interpolation problem

In animation the following problem arises. Let the unit quaternion space be denoted by $\mathcal{U}$. To model the gradual change of the orientation of an object in E^3 over a time interval, first the orientations of the object at a series of key-frames, called ***key-frame orientations***, are represented by a sequence of quaternions $\{q_i\} \subset \mathcal{U}$, which are called *key-frame quaternions*. Then a continuous curve in $\mathcal{U}$ is sought to interpolate $\{q_i\}$. Finally, a sequence of unit quaternions on the interpolating curve, which are called *in-between* quaternions and are usually denser than $\{q_i\}$, are computed to represent orientations of the object in a number of consecutive frames between the key-frames, called *in-between frames*. Because $\mathcal{U}$ can be identified with the unit sphere $S^3 \subset E^4$, the above problem is equivalent to designing a curve to interpolate a sequence of data points $\{q_i\}$ on S^3.

There have been several solutions proposed in the literature to solve the above problem. The simplest solution is to use the so called *Slerps* on S^3 to connect successive points q_i and q_{i+1}, where the Slerp is a minor arc on the great circle of S^3 that passes through q_i and q_{i+1}. This solution merely provides a G^0 interpolating spline curve, i.e. the curve is continuous. The evaluation of Slerps is discussed in [Sho85, Sho87].

Usually, a smooth interpolating curve, which is understood to be unit tangent vector continuous, or G^1 continuous, is desired in applications, because a kink on the interpolating curve means an abrupt change of the rotation axis. Note that by a simple reparameterization a G^1 curve can be made to be C^1. Now the problem is to design a G^1 interpolating curve in $\mathcal{U}$ while minimizing the cost of computing in-between quaternions. Shoemake [Sho85] proposes a spherical version of the cubic Bézier curve using the deCasteljau recursive construction, replacing the six linear interpolations by six Slerps. The resulting curve is known to have only an exponential parametric expression, and an in-between quaternion is computed by evaluating six Slerps. The cost of computing one in-between quaternion is more than 60 multiplications. Later, Shoemake [Sho87] gives an analogue of Boehm's quadrangle construction of cubic curves, called *Squad*, which requires three Slerps to evaluate a point on the curve. In [Ple89] spherical analogues of the cubic cardinal spline and the tensioned B-spline curve are used. These curves are defined by subdivision procedures and no explicit expressions have been obtained. In these schemes the cost of computing one in-between quaternion is about the same as for the Squad. Common features of the above schemes are that the curve is constructed directly on S^3 and that the computation of in-between quaternions is based on Slerps, which are relatively expensive to compute.

A different approach is taken in [GeR91], in which the parametric Hermite cubic interpolant is used to interpolate two points and two end tangents on S^3, and then the interpolant is normalized onto the sphere S^3. The advantage of this approach is that virtually any known interpolating curve scheme can be used to do the interpolation and then followed by the necessary normalization, thus providing more flexibility in controlling the shape of the curve or achieving higher continuity. The normalization step requires an extra square root and 8 extra multiplications/divisions in computing each in-between quaternion.

If points on the interpolant are obtained using forward differencing, this method is expected to provide a considerable time improvement over the approaches based on Slerps.

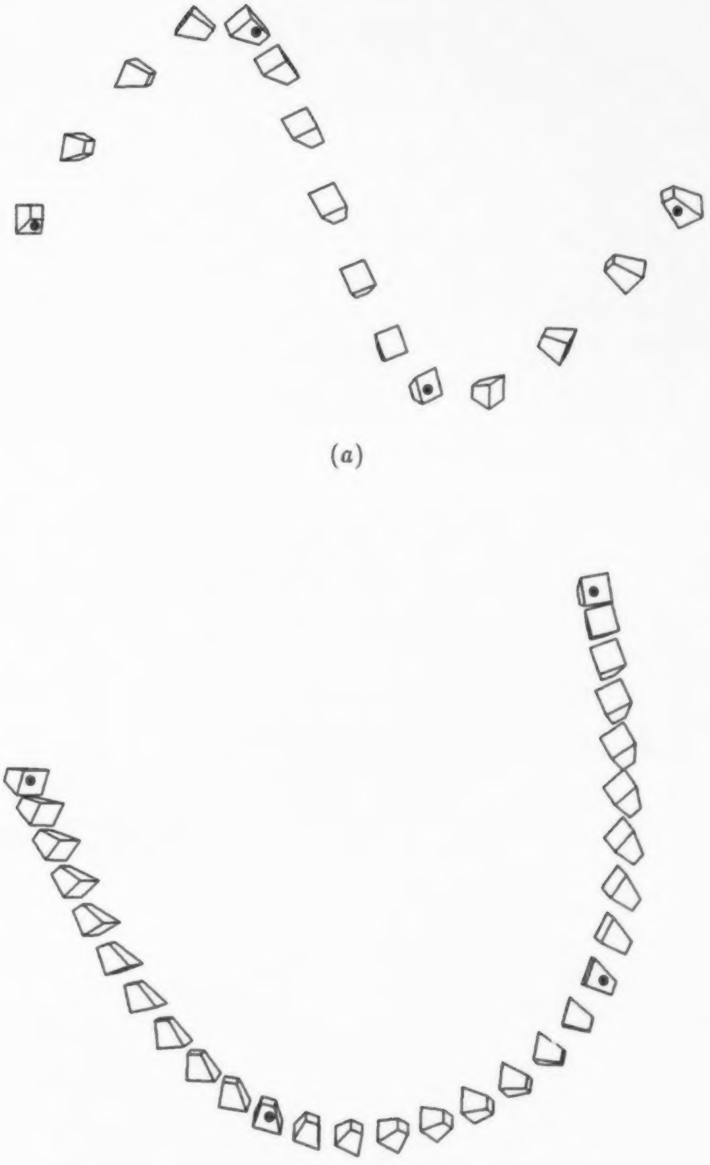
6.4.3 Spherical biarc solution

Now we explain how spherical biarcs are used to interpolate key-frame quaternions, thus giving a smooth interpolation of orientations. The spherical biarcs given in the Bézier form are in general only G^1 continuous. In our algorithm the arclength parameterization is used to make a spherical biarc C^1 continuous, and the difference method introduced in Chapter 2 is then used to compute inbetween quaternions on it. Simple examples are given to show the effects.

Initially, the positions (determined by translations) and orientations (determined by rotations) of an object are specified in a number of key-frames. It is required to interpolate these positions and orientations so that the object can smoothly change its position and orientation between the key-frames. In our examples the positions in key-frames are interpolated by a cubic B-spline curve. Key-frame orientations are first converted to a sequence $\{q_i\}_{i=1}^{n}$ of unit quaternions. As there are two antipodal unit quaternions on $S^3 \subset E^4$ corresponding to the same orientation, after q_1 has been determined, the subsequent quaternions are chosen in such a way that among the two candidates, q_{i+1} is the one that has the shortest distance to q_i on S^3, $i \geq 1$; a tie is broken arbitrarily. Then Algorithm 6.3.1 is used to generate a spline curve interpolating $\{q_i\}_{i=1}^{n}$, and a required number of in-between quaternions on the curve are computed using the difference method (2.12). Finally the in-between quaternions are converted to rotation matrices that determine the consecutive orientations of the object in in-between frames.

In Fig. 6.4.1 two examples of orientation interpolation by spherical biarcs are shown. The object is a frustum of a pyramid placed in space and two different paths for positions are used in Fig. 6.4.1 (*a*) and (*b*). The key-frames are marked by •.

In a sophisticated animation system the sampling of points on the cubic spline curve interpolating key-frames positions and in-between quaternions on the spherical biarc spline curve interpolating orientations should be determined by some animation parameters. The sampling is simplified in the above examples, in which the number of points sampled between two key-frames is approximately proportional to the distance between the key-frame positions. Moreover, the in-between quaternions on a single circular arc of the spherical biarc spline are assumed to be equidistant, which makes it possible to apply the difference method to compute the in-between quaternions.



(a)

(b)

Figure 6.4.1 Orientation interpolation: Two examples of interpolating object orientations by spherical biarcs. The key-frames are marked with •.

6.4.4 Comparisons

Now we compare all the methods mentioned so far. These include the spherical analogue of the cubic Bézier curve [Sho85], the Squad [Sho87], the spherical analogues of the cubic cardinal spline and the tensioned B-spline [Ple89], the normalized Hermite interpolant [GeR91], and the spherical biarc. All these methods lead to G^1 interpolating splines on S^3 and provide local control. The main difference is in the efficiency of computing in-between quaternions.

In the spherical analogue of the cubic Bézier curve an in-between quaternion is computed by evaluating six Slerps. According to [Sho87], three of these six Slerps are *static*, the other three are *dynamic*, where in a static Slerp the two end points are fixed and in a dynamic Slerp the two end points vary; 8 multiplications are needed to compute a point on a static Slerp and 15 multiplications are needed for a dynamic Slerp. So the total cost for computing one in-between quaternion in this scheme is about 69 multiplications and several table look-ups for the sin and arccos functions. The Squad requires two static Slerps and one dynamic Slerp to evaluate a point on the curve, thus reducing the cost of computing one in-between quaternion to about 31 multiplications and several table look-ups for trigonometric functions.

In the subdivision scheme for the spherical analogues of the cubic cardinal spline and the tensioned B-spline, two *Smids* and a dynamic Slerp are needed to compute one in-between quaternion, where the Smid is a special form of the Slerp which can be computed using 8 multiplications/divisions and a square root. So the cost of computing one in-between quaternion in this scheme is at least 31 multiplications/divisions and 2 square roots, which is about the same as the Squad.

In the normalized Hermite interpolant, on the average the forward differencing used to generate points on the unnormalized curve needs 12 additions for each point, and the subsequent normalization step for each point requires a square root and 8 multiplications/divisions.

The spherical biarc scheme needs 4 multiplications and 8 additions/subtractions. So if in an application only G^1 continuity is needed, in view of efficiency the spherical biarc is an obvious choice.

Observe that the cost of computing a static Slerp can be reduced to 4 multiplications per point using the difference method, since it can be evaluated at equally spaced parameter values. So the Squad curve can be made more efficient, requiring only 23 multiplications per in-between quaternion.

It is natural to ask whether there are other simple curves for doing G^1 interpolation in the unit quaternion space. Next to the quadratic curve, one may consider using the cubic curve. However, it is easy to see that no proper real cubic curve can be contained in the sphere $S^3 \subset E^4$, because a real cubic curve is unbounded, but S^3 is bounded. By analytic extension, the same is true of a finite piece of a cubic curve. Therefore, rational cubic curves can not be used for interpolation in unit quaternion space. More research is needed to study higher degree rational curves or other simple curve schemes for interpolation in the unit quaternion space with G^1 or higher continuity which allow efficient evaluation of in-between quaternions. Recently, a very interesting construction of rational curves on $S^2 \subset E^3$ is proposed in [HoS91]. There should be little difficulty in generalizing this construction to high dimensional spheres.

6.5 Summary

From the viewpoint of automatic design, in this chapter we have considered how to choose an appropriate spherical biarc interpolating given data D from the infinitely many available. It is found that the biarc with its two arcs having chords of equal length is a reasonable choice, and a simple formula for computing this biarc is derived.

The spherical biarc has several applications. First of all, it is a simple curve scheme for modeling features on a sphere. We have shown by examples the application of spherical biarcs to orientation modeling, where biarcs are used to interpolate points in the unit quaternion space.

Another application of the theory of spherical biarcs is to study space biarcs in Euclidean space. Recently, space biarcs in E^3 have been found useful in surface modeling using cyclides [Sha87, NuM88]. In the present investigations of space biarcs, including plane biarcs (space biarcs in E^2), mainly geometric methods are used. This approach is intuitive and has been very fruitful in deriving useful optimal plane biarcs in practical cases [SuL89]. But it has several drawbacks: it is not suitable for tackling the general case, consequently, degenerate cases are not clearly distinguished; and it does not apply to space biarcs in higher dimensional spaces. The above problems are easily solved with the theory of spherical biarcs. As stereographic projection is circle preserving, this map establishes a one-to-one correspondence between all spherical biarcs on $S^{d-1} \subset E^d$ and space biarcs in E^{d-1}. Therefore several properties of the spherical biarcs on S^{d-1} obtained algebraically in Chapter 5 and Chapter 6 can be transformed readily to space biarcs in E^{d-1}, $d \geq 3$.

Chapter 7

Concluding Remarks

7.1 Summary

In this thesis four topics concerning the theory and applications of conics and quadrics in computer graphics and geometric modeling have been studied. They are (1) the difference method for rendering conic arcs; (2) quadratic parameterizations of quadrics in E^3; (3) rational quadratic spline curve interpolation on a quadric in E^d, $d \geq 3$; (4) spherical biarc interpolation on a sphere in E^d, $d \geq 3$.

The difference method is the most efficient method for computing an inscribed polygon of a conic arc known so far. We have explained the theory of this method and studied its properties and some implementation issues. This method has been used successfully for drawing elliptic arcs using a microprocessor controlled plotter [Wan86].

In geometric modeling the rational parameterization of quadric surfaces in E^3 provides an alternate representation which is useful in many applications. There have been a number of papers published on this topic [FPW88, Sed85, AbB88]. Our contribution is to give a classification of all faithful rational quadratic parameterizations of a quadric. It is shown that (1) by projection from a center at a point on an irreducible quadric $S \subset E^3$, one can get a family of rational quadratic parameterizations of S, and all faithful rational quadratic parameterizations of S can be generated in this way; (2) two faithful rational quadratic parameterizations of S are related by a rational linear reparameterization if they have the same projection center, and related by a rational quadratic reparameterization if they have different projection centers; (3) any faithful rational quadratic parameterization of S has a rational linear inversion formula.

We then used the technique of parameterizing by projection to obtain a rational triangular Bézier representation of a triangular patch on a quadric S. It is shown that if a triangular patch satisfies certain practical restrictions and its boundaries are rational curve segments of degree at most n, then it can be represented as a rational triangular Bézier surface of degree at most $2n$. In particular, if there are three planes each containing a side of a triangular patch on S and the three planes intersect at a point on S outside the patch, then the patch can represented as a rational quadratic triangular Bézier patch.

The study of the rational quadratic spline curve interpolation on a proper quadric $S \subset E^d$, $d \geq 3$, is mainly a theoretical investigation. It is motivated by the application of smooth curve interpolation on a sphere in computer graphics [Sho85]. Given a point sequence $\{X_i\}_{i=1}^{n}$ on a same component of S, it is shown that a necessary condition for the existence of a G^1 rational quadratic spline curve interpolating $\{X_i\}_{i=1}^{n}$ is that all segments $\{\overline{X_i X_{i+1}}\}_{i=1}^{n-1}$ are on the same side of S. Also we show that for any point sequence on a sphere in E^d there exists a family of rational quadratic interpolating spline curves with one free parameter. As any rational quadratic interpolating spline curve on S, when one exists, is not locally controllable, we then turned to study biarcs to solve the Hermite interpolation problem on S, i.e. two points and their associated tangents on S are interpolated by a pair of smoothly joining rational quadratic Bézier curves. Only biarcs with their two rational Bézier curves having positive weights are considered and a necessary condition for their existence is given.

The theory of spherical biarc interpolation is developed in Chapter 5. This is the main contribution of this thesis. The spherical biarc provides a simple solution to the smooth curve interpolation problem on a sphere. Here spherical biarcs, which are curves composed of two smoothly joining circular arcs, are considered for interpolating two points X_0, X_1 and two tangent directions T_0, T_1 at the two points on a sphere in E^d, $d \geq 3$. Two kinds of data $D = \{X_0, T_0, X_1, T_1\}$ are distinguished: regular data, which is generic, and singular data. The existence and properties of spherical biarcs for both regular and singular data are studied in detail. The main conclusions are (1) spherical biarcs exist for any data and are easy to construct; (2) the family of spherical biarcs interpolating data D has one free parameter when D is regular, and $d - 2$ free parameters when D is singular.

In Chapter 6 the shape control of spherical biarcs is considered. Among the infinitely many biarcs interpolating given data D, by analysis and experiments we concluded that the biarc with its two arcs having chords of equal length is a satisfactory choice. Finally, the application of spherical biarc interpolation of unit quaternions for modeling orientation is illustrated by examples. The spherical biarc solution compares favorably with other existing methods in terms of efficiency and simplicity.

7.2 Open problems

Several open problems arise in our research.

(1) Identify which commonly used curves in CAGD can be generated efficiently by a difference equation. We have shown in Subsection 2.5.3 that some rational cubic space curves can not be generated by a fourth order difference equation.

(2) Efficiently change the step size of the difference method so that it can generate adjacent pixels on a conic directly in a raster device.

(3) Classify all rational quadratic parameterizations on a quadric in E^3, including unfaithful ones.

(4) More research is needed on shape control for spherical biarcs. One problem is how to provide a theoretical foundation for the empirical results in Chapter 6. The choice of biarc favored in our research is the equal chord biarc. This choice has the shortcoming that for some poorly behaved data the chosen biarc may contain a major circular arc. On the other hand, it has been shown that for regular data, which is generic, there always exist interpolating spherical biarcs that contain only minor arcs. Therefore it is hoped that one can find an explicit formula or an efficient procedure which gives an interpolating spherical biarc that contains two minor arcs, and the difference of the radii of its two arcs is minimal.

(5) Apply the theory of spherical biarcs to study space biarcs. This problem is briefly addressed in Section 6.5.

(6) Use rational curves to construct interpolating splines of higher order continuity on a sphere. Research in this direction has been done in [HoS91] in which rational curves on $S^2 \subset E^3$ are constructed. Another possible approach is to use stereographic projection to map a rational spline curve in E^d, which is relatively easy to construct, onto the sphere S^{d-1}. A problem with this approach is controlling the shape of the image curve on S^{d-1}.

Bibliography

[Alt86] Altmann, S.L. (1986), *Rotations, Quaternions, and Double Groups*, Clarendon Press, Oxford.

[BBB87] Bartels, R.H., J.C. Beatty, and B.A. Barsky (1987), *An Introduction to Splines for Use in Computer Graphics & Geometric Modeling*, Morgan Kaufmann Publishers, Los Altos, CA.

[Bez72] Bezier, P. (1972), *Numerical Control*, John Wiley & Sons, London.

[Bol75] Bolton, K.M. (1975), Biarc curves, *Computer Aided Design*, vol. 7, no. 2, pp. 89-92.

[Boo79] Bookstein, F.L. (1979), Fitting conic sections to scattered data, *Comput. Graph. Image Process.*, vol. 9, pp. 56-71.

[Chi90] Chionh, E.W. (1990), *Base Points, Resultants, and the Implicit Representation of Rational Surfaces*, Ph.D Thesis, Department of Computer Science, University of Waterloo.

[ChR92] Chionh, E.W. and R. Goldman (1992), Degree, multiplicity, and inversion formulas for rational surfaces using u-resultants, *Computer Aided Geometric Design*, vol. 9, pp. 93-108.

[Coh70] Cohen, D. (1970), On linear difference curves, *Proceedings of International Symposium on Computer Graphics 70*, Brunel University, Uxbridge, UK.

[Dah89] Dahmen, W. (1989) Smooth piecewise quadric surface, *Mathematical Methods in Computer Aided Geometric Design*, edited by T. Lyche and L. Schumaker, Academic Press, Inc., pp 181-193.

[Far88] Farin, G. (1988), *Curves and Surfaces for CAGD: A Practical Guide*, Academic Press Inc., San Diego.

[FPW88] Farin, G., B. Piper, and A. Worsey (1988), The octant of a sphere as a nondegenerate triangular Bézier patch, *Computer Aided Geometric Design*, vol. 4, no. 4, pp. 329-332.

[Far89] Farin, G. (1989), Curvature continuity and offsets for piecewise conics, *ACM Trans. on Graph.*, vol. 8, no. 2, pp. 89-99.

[Fau79] Faux, I.D. and M.J. Pratt (1979), *Computational Geometry for Design and Manufacture*, Ellis Horwood Limited, England.

[GeR91] Ge, Q.J. and B. Ravani (1991), Computational geometry and mechanical design synthesis, *13th IMACS World Congress on Computation and Applied Math.*, Dublin, Ireland, pp. 1013-1015.

[GSA84] Goldman, R.N., T.W. Sederberg and D.C. Anderson (1984), Vector elimination: A technique for the implicitization, inversion and intersection of planar parametric rational polynomial curves, *Computer Aided Geometric Design* vol. 1, pp. 327-356.

[GoV89] Golub, G.H. and C.F. Van Loan (1989), *Matrix Computations*, 2nd edition, The Johns Hopkins University Press.

[HoS91] Hoschek, J. and G. Seeman (1991), Spherical splines, Preprint.

[JoW90] Joe B. and W. Wang, Reparameterization of rational Bézier curves and triangular Bézier surfaces, submitted for publication, Aug. 1990.

[Lan85] Lancaster, P. (1985), *The Theory of Matrices*, Academic Press, Orlando, Florida.

[Lee87] Lee, E.T.Y. (1987), The rational Bézier representation for conics, *Geometric Modeling: Algorithms and New Trends*, ed. G.Farin, SIAM, Philadelphia, pp. 3-19.

[LoW90] Lodha, S. and J. Warren (1990), Bézier representation for quadric surface patches, *Computer Aided Design*, vol. 22, no. 9, pp. 574-579.

[NuM88] Nutbourne, A.W. and R.R. Martin (1988), *Differential Geometry Applied to Curve and Surface Design, Vol 1: Foundations,* Ellis Horwood Limited, England.

[O'Ne75] O'Neil, P.V. (1975), *Advanced Calculus*, Macmillian Publishing Co., New York.

[Pat86] Patterson, R.R. (1986), Projective transformations of the parameter of a Bernstein-Bézier curve, *ACM Trans. on Graph.*, vol. 4, no. 4, pp. 276-290.

[Pat88] Patterson, R.R. (1988), Parametric cubics as algebraic curves, *Computer Aided Geometric Design* vol. 5, pp. 139-159.

[Pav83] Pavlidis, T. (1983), Curve fitting with conic splines, *ACM Trans. on Graph.*, vol. 2, no. 1, pp. 1-31.

[Ple89] Pletinckx, D. (1989), Quaternion calculus as a basic tool in computer graphics, *The Visual Computer*, vol. 5, pp. 2-13.

[Pra85] Pratt, V. (1985), Techniques for conic splines, *Proceedings of SIGGRAPH 85*, pp. 151-159.

[Pro83] Prosser, C.J. and A.C. Kilgour (1983), An integer method for the graphical output of conic sections, *ACM Trans. on Graphics*, vol. 2, no. 3, pp. 182-191.

[Sab76] Sabin, M. (1976), *The use of piecewise forms for the numerical description of shape*, Ph.D. Thesis, Hungarian Academy of Sciences.

[Sed85] Sederberg, T.W. (1985), Steiner surface patches, *IEEE Computer Graphics and Applications*, vol. 5, no. 5, pp. 23-36.

[Sed86] Sederberg, T.W. (1986), Improperly parameterized rational curves, *Computer Aided Geometric Design* vol. 3, pp. 67-75.

[SeK52] Semple, J.G. and G.T. Kneebone (1952), *Algebraic Projective Geometry*, Oxford University Press, London.

[Sha87] Sharrock, T.J. (1987), Biarcs in three dimensions, *The Mathematics of Surfaces II*, ed. R.R. Martin, Clarenden Press, Oxford, pp. 395-411.

[Sho85] Shoemake, K. (1985), Animating rotation with quaternion curves, *Proceedings of SIGGRAPH 85*, pp. 245-254.

[Sho87] Shoemake, K. (1987), Quaternion calculus and fast animation, *SIGGRAPH 87 Course Notes #10: Computer Animation: 3D Motion Specification and Control*, pp. 101-121.

[Smi71] Smith, L.B. (1971), Drawing ellipses, hyperbolas or parabolas with a fixed number of points and maximum area, *Computer Journal*, vol. 14, no. 1, pp. 81-86.

[Som51] Sommerville, D.M.Y. (1951), *Analytic Geometry of Three Dimensions*, Cambridge University Press, England.

[SuL89] Su, B.Q. and D.Y. Liu (1989), *Computational Geometry: Curve and Surface Modeling*, Academic Press, San Diego.

[Wal50] Walker, R.J. (1950), *Algebraic Curves*, Princeton University Press.

[War92] Warren, J. (1992), Creating multisided rational Bézier surfaces using base points, *ACM Transactions on Graphics*, vol. 11, no. 2, pp. 127-139.

[Wan86] Wang, W. (1986), *The Design of Controller for High-Speed Intelligent Poltter*, Master's Thesis, Department of Computer Science, Shandong University, China.

[WaW89a] Wang, W. and C.Y. Wang (1989), Difference method for generation of circular arcs and ellipses, *Computer-Aided Design*, vol. 21, no. 1, pp. 33-36.

[WaW89b] Wang, W. and C.Y. Wang (1989), A new method for generating conic sections, *Journal of Comput. Sci. & Tech.*, vol. 4, no. 2, Sinica Academia, pp. 155-162.

[WJW92] Wang, W., B. Joe and C.Y. Wang (1992), On the difference method for drawing conic arcs, to appear in *CVGIP: Graphical Models and Image Processing.*

[Wil65] Wilkinson, J.H. (1965), *The Algebraic Eigenvalue Problem*, Clarendon Press, Oxford.

[Xu65] Xu, Y.C. (1965), *Introduction to Algebra*, Shanghai Science and Technology Press, in Chinese.